5-2018

The Genetic Basis of Social Behaviors in Yeast: An Investigation into FLO11

Zachary Oppler

The Genetic Basis of Social Behaviors in Yeast: An Investigation into *FLO11*

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor Science in Biology from
The College of William and Mary

by

Zachary Joseph Oppler

Accepted for Honors

Dr. Helen Murphy, Director

Dr. Joshua Puzey

Dr. Jelena Pantel

Dr. Leah Shaw

Williamsburg, VA
May 4, 2018

Table of Contents

Acknowledgments

Before I say anything else, I would like to thank Dr. Helen Murphy for being the best advisor that I could have possibly asked for. Her dedication to her students is truly remarkable and she inspires me to be a better scientist each and every day that I step into the lab. I am extremely grateful for everything that she has taught me and I hope to carry the lessons I've learned from her with me for the rest of my career.

I would also like to say a big thank you to all of the past and present members of the Murphy Lab that have made my time working on this project so enjoyable. In particular, I would like to thank Danting Jiang for helping me with many of the computational techniques that I used in this project, and Meadow Parish for all of her assistance with the wet lab aspects of this project.

I would like to thank all my members of my Honors committee, including Dr. Joshua Puzey, Dr. Jelena Pantel, and Dr. Leah Shaw, for all of their time and support. I really appreciate all of the feedback that you provided. In addition, I would like to say thank you to Dr. Doug Young for his assistance with the protein structure figures and to Lidia Epp from the Core Lab for assisting us with some of the molecular techniques that were used in this project.

I would like to thank the Jeffress Memorial Trust Awards Program as well as the William and Mary Charles Center for funding.

Finally, I would like to thank all of my friends and family for all of the support they have provided me throughout the duration of this project.

Abstract

Microorganisms were once assumed to live as solitary cells interacting with just their physical environment; however, like most multicellular organisms, they have been found to engage in complex social behaviors that play an important role in their ability to survive and reproduce. Unlike competitive behaviors, such as chemical warfare and antibiotic production, cooperative behaviors have been more challenging to explain from an evolutionary perspective. In multicellular organisms, most cooperative behaviors can be explained by kin selection and kin recognition. In clonally growing microbes, cooperative behaviors involving cells adhering to one another generally rely on "kind" recognition, whereby a single locus or trait, referred to as a greenbeard, is enough to signal cooperation.

Microbes frequently rely on membrane-associated proteins with variable extracellular domains to recognize one another. In the budding yeast, *Saccharomyces cerevisiae*, *FLO11*, a highly regulated gene, encodes a cell-surface adhesin that allows individual cells to attach to one another and plays a crucial role in social phenotypes, such as the formation of biofilms and other structured communities, which are critical to survival during rapid environmental changes.

To characterize the amount of genetic variation at *FLO11*, its regulatory and coding regions were amplified and sequenced in 78 environmental isolates that vary in their social phenotypes; *de novo* assemblies of the locus were generated. Population genetic analyses suggest that the precise regions implicated in cell-cell adhesion exhibit a signature of positive selection, while the rest of the gene is under purifying selection. Furthermore, a region of the upstream regulatory sequence exhibits a signal of balancing selection.

Phenotypic assays demonstrate that different natural *FLO11* alleles generate diverse biofilm architectures in an otherwise constant genetic background. These assays also shed light on the complex role that natural regulatory variation and variegated expression patterns may play in the outcome of inter-clonal competitions.

Our results suggest a "shades of greenbeard" system in which Flo11p preferentially adheres to like kinds. Unlike in motile microbes where cheater avoidance is likely driving the evolution of recognition, homophillic binding of Flo11p may be selected during competition among clones. Thus, the interplay between inter-clone competition and intra-clone cooperation in spatially structured microbial communities can potentially lead to recognition systems.

Introduction

Microbes as Social Beings

Microorganisms frequently cooperate with one another and engage in complex social behaviors that play a critical role in their survival and reproduction. Microbes have been observed to engage in cooperative activities, such as foraging, dispersal, and various forms of communication, including signaling and quorum sensing (Crespi, 2001; Keller and Surette, 2006, West *et al*., 2007). Cooperation in microbes frequently involves cell-cell adhesion or the production of public goods. Over the last few decades, researchers have gained an understanding of the genetic basis of these cooperative behaviors as well as the underlying molecular mechanisms that are involved in their regulation (West *et al*., 2006).

The ubiquity of cooperative behaviors in microbes is hypothesized to result from the need to carry out many processes in the extracellular space that larger organisms are able to keep private (Strassmann *et al*., 2011). Many social phenotypes involve the secretion of various molecules, like enzymes, adhesive polymers, and antibiotics (West *et al*., 2007; Nadell *et al*., 2016). Once these molecules are produced and secreted, they can be treated as "public goods" to be utilized by other individuals in the community.

Cooperation is difficult to explain evolutionarily, because it is metabolically costly for individuals to produce public-goods molecules, yet, the majority of the benefits are gained by individuals other than the producer (Allen *et al*., 2013). Non-cooperators can take advantage of cooperators by using up the public goods without producing any in return. Because these individuals do not have to bear the metabolic cost of production, they will have a major fitness advantage over the producers; the subsequent rise in frequency of the "cheaters" can ultimately lead to a collapse of the cooperative system.

A common and ancient form of microbial cooperation is the formation of biofilms, resilient communities of cells that adhere to one another and are protected against environmental threats, like antimicrobial therapies and host defense systems (Reynolds and Fink, 2001; Nett and Andes, 2015; Váchová *et al*., 2011). Because biofilms are particularly good at growing on abiotic surfaces, they pose a major health risk in settings where they can adhere to plastic medical devices and implants, and gain easy access to patients' bloodstreams (Verstrepen and Klis, 2006). Individual cells residing in the biofilm cooperate to produce a protective

extracellular matrix as well as other public goods that are beneficial to the community as a whole (West, 2006).

Cooperation and Kin Recognition

Social behaviors are common in nature, and researchers have described the conditions under which cooperation can evolve and be stable. Most cases of cooperation can be readily explained by kin selection; cooperation can evolve if individuals are able to preferentially direct their help to individuals who are likely to share their own genetic material (Smith, 1964; Hamilton, 1964). By providing assistance to relatives, an organism can indirectly increase its own genetic representation in the following generation, and in this way, genes for cooperation can spread. As described by Hamilton's rule, the benefits gained through this increase in inclusive fitness must outweigh the cost of the cooperative behavior in order for cooperation to evolve.

This ability to recognize kin turns out to be much more important in well-mixed populations in which unrelated individuals have a high probability of coming into contact with one another. Microbes growing in liquid culture, or within communities that do not have highly structured populations, are much more reliant on kinship cues so that they can preferentially direct their cooperative behaviors towards relatives (Queller, 2011). On the other hand, non-motile organisms growing in a spatially structured community do not need to discriminate between kin and non-kin because there is a high likelihood that all close-by individuals are clones, and thus the benefits of helping are high. Mathematical models have shown that spatially expanding populations tend to lose genetic diversity at the frontiers due to genetic bottlenecks, thus promoting the evolution of cooperation (Van Dyken *et al*., 2013).

Cooperative behaviors in microbes, particularly those that occur through cell-cell adhesion, tend to be dependent on "kind" recognition, whereby individuals recognize each other on the basis of one single trait rather than on overall genetic relatedness (Strassmann *et al*., 2011; Queller, 2011). Individuals provide preferential treatment based on the presence or absence of the single trait, regardless of the relatedness between other loci throughout the genome.

One way that kind selection has been theorized to work is through a particular type of gene that has been termed a "greenbeard" gene. It ultimately works by producing three effects: it produces a trait, it recognizes that trait in others, and it allows for preferential treatment of others

who possess that trait (Dawkins, 1976). The greenbeard gene could thus allow individuals who possess it to identify one another and then direct helping behavior towards these specific individuals. This can occur between both relatives and non-relatives and cooperation between bearers of the greenbeard gene is able to persist. The greenbeard restricts cooperation to individuals who are not cheaters, especially if the trait is costly or difficult to produce.

Social Yeast

Cooperative behaviors have not only been found in bacterial species, but eukaryotic species as well, including the budding yeast, *Saccharomyces cerevisiae*. Humans have been using this yeast to produce bread and fermented beverages for thousands of years (Legras *et al*., 2007). Over the past century, *S. cerevisiae* has been used as a model organism in biomedical research and has been instrumental in studying virtually every aspect of molecular biology and genetics (Liti, 2015). It is a relatively simple eukaryotic organism with an extremely quick cell cycle and is easily manipulated with genetic techniques.

It is only within the last decade or so that researchers have realized that *S. cerevisiae* is capable of engaging in a wide array of complex social behaviors (Figure 1), including complex biofilm colony and/or mat formation, invasive growth, plastic adherence, and flocculation (Reynolds and Fink, 2001; Hope and Dunham, 2014; Van Mulders *et al*., 2009).

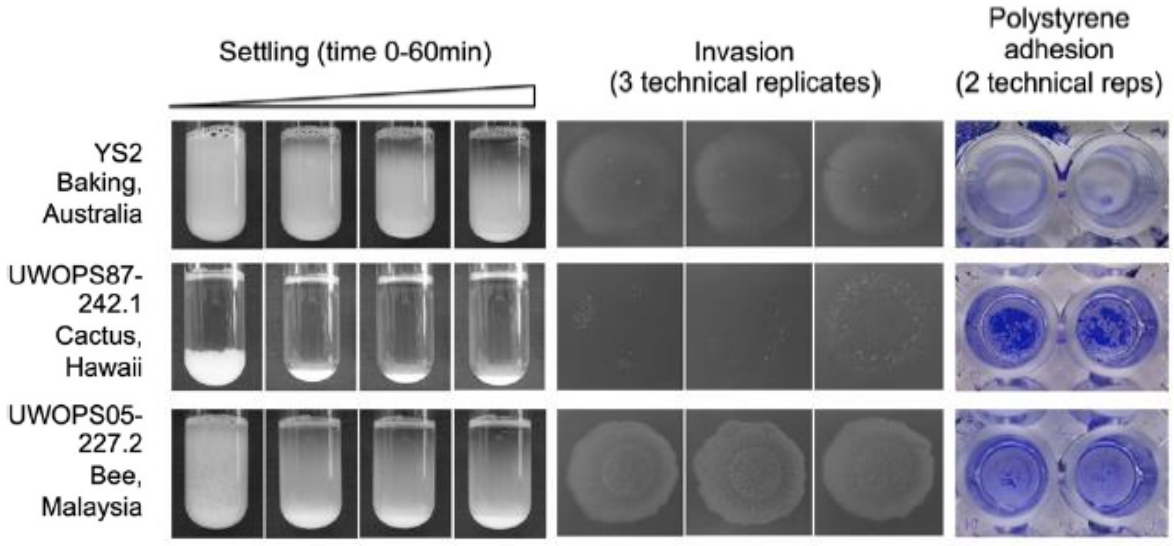


Figure 1: Natural isolates of *S. cerevisiae* express a diverse array of social phenotypes; flocculation, invasive growth, and plastic adherence are shown here. (Hope and Dunham, 2014)

A diverse array of morphologies of yeast biofilm colonies (Figure 2) have been observed (Granek and Magwene, 2010; Hope and Dunham, 2014), and the genetic basis underlying their formation is beginning to be understood (Tan *et al*., 2013; Voordeckers *et al*., 2012, Granek *et al*., 2013). The relatively recent focus on social behaviors stems from the fact that most of the domesticated strains that are used for research in the lab have recessive mutations that prevent the expression of several social phenotypes and preclude the formation of complex colonies (Voordeckers *et al*., 2012). It is probably the case that when geneticists first began working with wild yeast in the lab, they had a difficult time conducting experiments with social yeast and selected for those in the planktonic form (Liu, 1996). It has been shown that the domestication of wild yeast that are brought into the lab lose their ability to form complex colonies and only show the smooth colony phenotype (Kuthan *et al*., 2003).

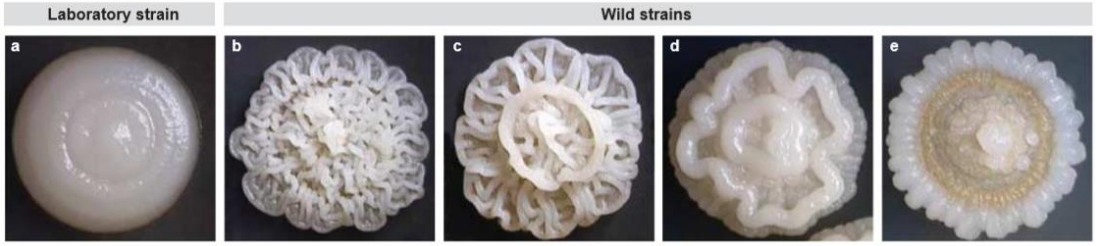


Figure 2: *S. cerevisiae* cultures grown in optimal laboratory conditions differ morphologically from those grown under harsher environmental conditions; lab strains tend to form "smooth" colonies while wild strains form "fluffy", structured colonies (Palková, 2004).

Yeast Adhesins as Modulators of Social Behavior

Many of the social phenotypes in microbes that require cell-cell adhesion are dependent on cell surface proteins known as adhesins. Adhesion proteins in yeast are typically characterized by a three-domain structure in which each domain performs a unique function (Dranginis *et al*., 2007). At the C-terminal end of the adhesin, a GPI-anchor attachment site allows the protein to become embedded into the cell surface. The middle portion of the adhesin is stalk-like and highly repetitive and is rich with serine and threonine residues. The length of this repetitive region correlates with the ability to adhere to nearby cells, likely a result of the increased protrusion of the N-domain away from the cell wall, which allows it to interact more easily with other cells (Zara *et al*., 2009). The N-terminal domain is responsible for substrate binding; many

of the adhesins have lectin-like binding domains, such that they adhere to sugar residues protruding from the surface of other cells (Verstrepen and Klis, 2006).

In the budding yeast *Saccharomyces cerevisiae*, these adhesins are encoded by a set of related genes that constitute the *FLO* family (Verstrepen and Klis, 2006). This group of cell adhesion proteins plays a critical role in the ability of yeast cells to adhere not only to each other but to abiotic surfaces as well (Váchová *et al*., 2011). The *FLO* genes are sensitive to the environment, enabling pseudohyphal growth in diploids and invasive growth in haploids in response to nutrient limitation (Lo and Dranginis, 1998). Yeast cells also display different phenotypes in response to changing temperatures (Soares, 2011) and pH levels (Barrales *et al*., 2008).

The *FLO* family is composed of five closely related genes: *FLO1, FLO5, FLO9, FLO10,* and *FLO11.* All of these genes with the exception of *FLO11* are located near telomeres, which along with the frequent recombination and slippage events involving tandem repeats accounts for the extensive variation present in these adhesins (Van Mulders *et al.*, 2010). New *FLO* alleles are constantly arising, resulting in a large reservoir of cell surface proteins that can be utilized by pathogenic yeast to avoid detection by host immune systems (Halme *et al*., 2004). Each of the *FLO* genes contributes differently to the expression of the various social phenotypes, but the majority of these phenotypes are dependent on *FLO11* in particular; *FLO11* expression is necessary for pseudohyphal and invasive growth (Figure 3) (Guo *et al*., 2000; Lo and Dranginis, 1998), adherence to plastics (Verstrepen *et al*., 2004), flor formation at the air-liquid surface (Zara *et al*., 2009), and mat and biofilm formation (Reynolds and Fink, 2001).

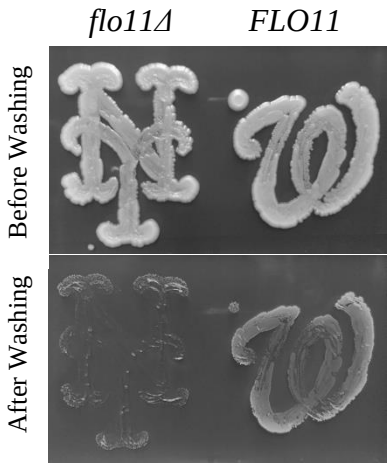


Figure 3: After growing for 5 days on YPD agar plates, only cells expressing *FLO11* are able to grow invasively, while cells lacking *FLO11* are washed away.

FLO11 Regulation

At 3.5 kb, the promoter of *FLO11* is considered to be one of the largest in the entire yeast genome, as well as one of the most highly regulated (Figure 4) (Octavio *et al*., 2009). Three well-known signaling cascades converge on this promoter, including the MAP kinase pathway,

the Ras-cAMP pathway, and a glucose repression pathway. Other pathways are also believed to play a role in adhesive phenotypes but are less well understood and may be involved in a more indirect way (Verstrepen and Klis, 2006). Analysis of the promoter revealed at least nine repression sites and four activation sites involved in *FLO11* expression, with at least one of these repression sites located over 2.5 kb upstream of the ORF (Rupp *et al*., 1999). Removal of a 111 bp portion of the promoter resulted in increased expression of *FLO11*, implicating the region as a repression site (Fidalgo *et al*., 2006).

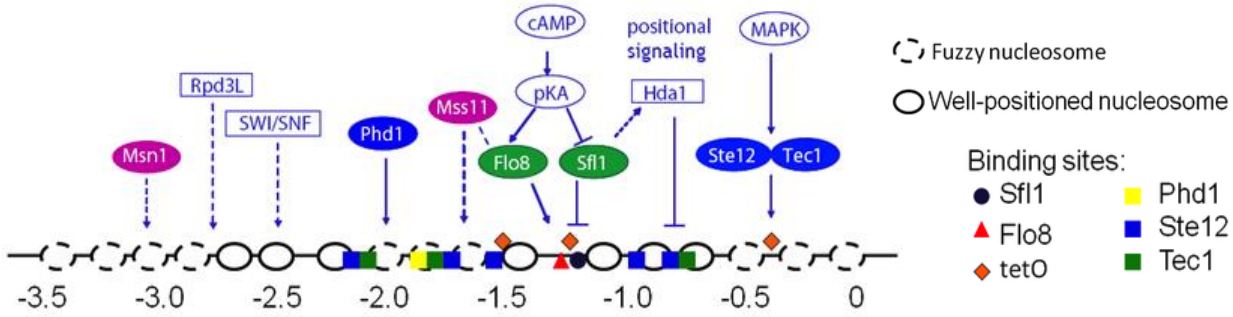


Figure 4: An abundance of regulatory factors from a range of signaling pathways converge at the *FLO11* promoter, signifying the importance of the decision regarding whether or not to engage in social behaviors (Octavio *et al*., 2009).

The three signaling pathways, along with chromatin remodelers, transcription factors, and two long non-coding RNAs— *ICR1* and *PWR1*— contribute to the variegated pattern of *FLO11* expression observed in clonal cell populations (Bumgarner *et al.,* 2012). *ICR1* is approximately 3.2 kb in length and is transcribed across almost the entire promoter region along the same direction as *FLO11*. *PWR1* is located on the opposite strand, and is approximately 1.2 kb in length (Bumgarner *et al*., 2009). A relative lack of conservation of the promoter region where *ICR1* and *PWR1* overlap serves as an indication that transcription itself rather than the DNA sequence is important for *FLO11* regulation. *ICR1* represses *FLO11* by a promoter occlusion mechanism, interfering with the binding of other key trans-acting factors. *PWR1* promotes *FLO11* expression by interfering with *ICR1* transcription (Gullerova and Proudfoot, 2011). Complete transcription of *ICR1* may reset the promoter to its basal state; subsequent competitive binding of the transcription factors Flo8 and Sfl1 determine whether *FLO11* will become activated or silenced (Bumgarner *et al*., 2012). Binding of Flo8 results in stabilization of the active state while binding of Sfl1 results in the silent state.

Within a population of clones, the promoter toggles between a silenced and competent state, causing *FLO11* to be turned "on" in one subset of individuals, but turned "off" in another subset of individuals (Octavio *et al*., 2009). This complex regulatory circuitry allows for a population to quickly generate phenotypic variation in response to rapid environmental changes; some cells can form filaments and grow invasively to forage for nutrients while other cells continue to float freely in search of more favorable environmental conditions (Bumgarner *et al*., 2012; Fidalgo *et al*., 2008). *FLO11* can also be regulated epigenetically, as daughter cells inherit the same promoter state as their parents due to the activity of the histone deacetylases Hda1p and Rpd3L (Halme *et al*., 2004, Octavio *et al*., 2009). The promoter state is inherited for several generations while environmental conditions are held constant, but the state is likely to switch in response to environmental fluctuations.

The differentiation of cells into adhesive and non-adhesive types may provide a competitive advantage over single-state populations. Differentiated populations were shown to have higher growth rates and better utilization of resources in their environment (Regenberg *et al*., 2016). Different combinations of regulators give rise to either fast or slow promoter fluctuations, which can contribute to the overall phenotypic heterogeneity of a population (Octavio *et al*., 2009). This allows cells to directly control the proportion of adhesive and non-adhesive cells within the population.

FLO11 A-Domain

Members of the *FLO* family of adhesins show a much higher degree of similarity to one another than *FLO11* does (Lo and Dranginis, 1996). Although Flo11p is very similar in its overall protein structure to the other adhesins in the family, its N-terminal domain does not contain the same mannose-binding structure that all of the other adhesins possess (Goossens and Willaert, 2012), and its nucleotide sequence is only about 30% similar to other genes in the *FLO* family (Kraushaar *et al*., 2015). Other proteins in the family confer social phenotypes through lectin-mediated adhesion; their N-terminals bind to sugar residues on adjacent cell walls, allowing cells to grab hold of one another (Nayyar *et al*., 2017). Flo11p does not demonstrate lectin-binding activity but rather confers social phenotypes through homotypic interactions that are dependent on the apical region of the N-terminal domain (Figure 5) (Kraushaar *et al*., 2015).

Rather than adhering to the cell surface of other cells, Flo11p binds to other Flo11p. It is likely that Flo11p's only adhesive target on adjacent cell walls is Flo11p (Douglas *et al*., 2007).

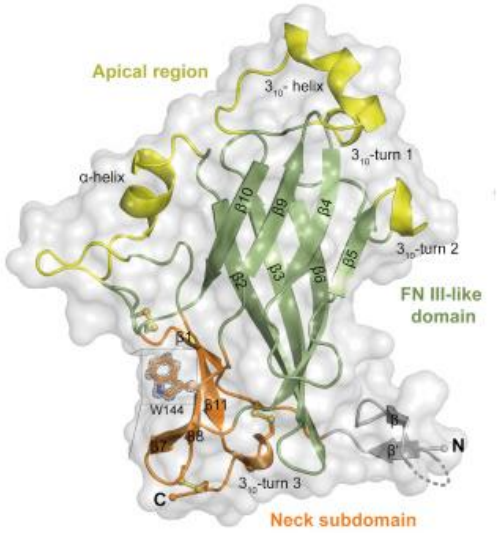


Figure 5: Molecular analysis of the N-domain has revealed its structural similarity to a fibronectin type III domain due to a pair of antiparallel beta sheets. The A domain contains an unusually large proportion of aromatic residues on either side of this beta sandwich, forming a ring structure which encompasses a large part of the apical region. The hydrophobicity conferred by these aromatic bands is critical for the binding ability of the protein (Kraushaar *et al*., 2015).

FLO11 has been implicated as a potential self-recognition gene and its protein product has been shown to bind preferentially to strain specific Flo11p. Cells are more likely to adhere to one another if they have the same or very similar *FLO11* allele (Barua *et al*., 2016). When plastic beads were coated with either of two allelic variants of Flo11p from two different strains, the cells bound preferentially to the beads sharing their *FLO11* allele (Douglas *et al*., 2007). Similarly when a *FLO11* allele was expressed in a non-native strain, the transformed cells preferentially adhered to the strain from which the *FLO11* allele was taken, indicating binding specificity for similar alleles.

Flo11p-Flo11p homotypic binding may act as a barrier against the incorporation of unrelated individuals into the complex social structures that are metabolically costly to produce. Anytime that a population of cooperators arises in nature, it is susceptible to cheaters who can benefit from public goods without having to contribute anything themselves (West, 2007). Because Flo11p binds to other Flo11p, it is possible that only cells that share a closely related *FLO11* allele will be incorporated into the social structure and benefit from the public goods. It is worth noting that this binding specificity took place regardless of similarity of the rest of the genetic background of the cells involved.

FLO11 B-Domain

As noted above, many genes encoding cell surface proteins contain highly repetitive regions that are frequently involved in recombination events (Verstrepen *et al*., 2005). The continual production of novel surface proteins contributes to virulence by increasing the diversity of cell surface antigens allowing cells to evade host immune systems (Nather and Munro, 2008). This source of variation can also be advantageous in terms of adjusting to fluctuations in environmental conditions.

The length of the *FLO11* ORF is highly polymorphic and the majority of this variation can be attributed to the expansion and contraction of the tandem repeats in the B-domain (Zara *et al*., 2009). A study of 20 strains found 13 alleles, each with a different number of a 36 bp tandem repeat in this middle domain. The presence of several different repetitive units throughout the region further complicates the process of aligning sequences from different strains.

This repetitive region is commonly thought of as a "spacer" that really only functions to allow the A-domain to protrude further away from the cell wall so that it can interact with other nearby cells (Dranginis *et al*., 2007). Flocculation has been shown to correlate highly with the number of repeat units, although extremely long alleles tend to show weaker phenotypes too (Fidalgo *et al*., 2008). Another study on fungal cell surface proteins revealed that not only do the number of repeats have an impact on overall protein function, but the different repetitive domains' relative positions to one another may have an additional impact on function as well (Huang *et al*., 2003). Some social phenotypes may depend more on the distribution of different repetitive units throughout the gene than the total number of repeats (Verstrepen *et al*., 2004).

FLO11 C-Domain

The C-terminal portion of *FLO11* encodes a glycosylphosphatidylinositol (GPI)-anchor signal sequence. The GPI anchor allows the protein to attach itself to the cell surface, protruding into the extracellular environment where it can adhere to adjacent cells and abiotic surfaces (Verstrepen and Klis, 2006). The addition of a GPI anchor is one of many posttranslational modifications cell surface proteins undergo while being transported through the secretory pathway. GPI-anchored proteins are involved in a wide variety of biological functions in eukaryotes (Paulick and Bertozzi, 2008).

The GPI signal sequence is composed of three domains, the most important of which is the residue at the attachment (or ω) site along with the following two amino acids in the sequence (Hamada *et al*., 1999). The GPI signal is cleaved from the rest of the protein at the ω site, which is where the GPI anchor becomes attached to the protein. The amino acid residues located near the ω site play a key role in determining whether the protein will become a part of the cell membrane or the cell wall (Dranginis *et al*., 2007). Investigation of these amino acids has revealed that the residue located two amino acids upstream of the attachment site is the most critical in determining protein localization; if the residue is a Y or N, as it is in Flo11p, the protein will be incorporated into the cell wall (Hamada *et al*., 1999). Other research has demonstrated the importance of other features within the protein, such as abundance of Ser/Thr residues in determining the localization of GPI-anchored proteins (Frieman and Cormack, 2004). In the case of Flo11p, upon arrival at the cell surface, the GPI anchor is cleaved off of the protein and the GPI remnant is covalently attached to the beta-1,6 glucans of the cell wall (Verstrepen *et al*., 2004).

Research Aims

Based on the complex network of regulatory control that converges on its promoter, the similar structure of its protein product to other proteins that are known to function in cell-recognition, and the critical role that it plays in the expression of social phenotypes, we hypothesized that *FLO11* serves as a kin-recognition mechanism in *Saccharomyces cerevisiae*. The research presented here was conducted to achieve the following three goals:

Aim 1: Characterize the amount of genetic variation at *FLO11*, and specifically, test the hypothesis that the "interacting" regions are under positive selection. The elevated rate of evolution and the repetitive nature of the locus have made it technically challenging to uncover the population genetic variation of this gene. To quantify the variation, *de novo* assemblies of the *FLO11* locus were generated for 78 strains. Sliding window and evolutionary analyses were conducted to examine the signatures of selection in the different regions of the gene.

Aim 2: Determine whether natural *FLO11* alleles produce different social phenotypes. First, an examination of social phenotypes, including invasive growth, mat formation, and biofilm formation, was conducted in order to characterize the variety of phenotypes expressed by different *FLO11* alleles in their natural backgrounds. Next, a further set of phenotyping was

conducted on a constant genetic background with different *FLO11* alleles in order to isolate their role in social phenotypes.

Aim 3: Determine whether the amount of non-synonymous variation at *FLO11* is different than other similar classes of genes in the same population (i.e., GPI-anchored proteins, cell membrane-associated proteins, etc.). In order to determine whether the level of variation was different in *FLO11* than in other similar genes, dN/dS analyses were conducted for 42 other genes.

Materials and Methods

Strains

The 78 strains of *S. cerevisiae* that were used in this analysis are listed in Table 1, and came from both publicly available and private collections (Strope *et al*., 2015; Liti *et al*., 2009; Sniegowski *et al*., 2009; Murphy and Zeyl, 2012). This collection is representative of a wide variety of ecological niches and includes both wild strains and clinical isolates that were collected from all over the world.

Phenotyping Assays

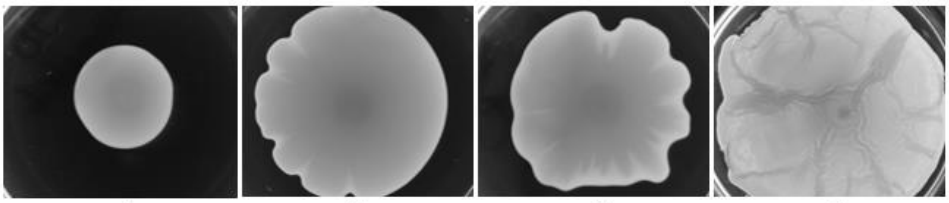
For all phenotyping assays, the panel of 78 strains was grown overnight in a 96-well plate containing 200 μL of standard yeast-peptone dextrose (YPD) liquid medium in each well. This overnight culture was used on the same day to inoculate mats on viscous medium, biofilm colonies on solid agar, and invasive colonies on solid agar. Four technical replicates were produced for each strain. The phenotyping assays were performed in collaboration with Meadow Parrish (2 replicates per person).

Mat formation

2 μL of each culture was spotted on 0.3% agar low-dextrose (LD) and 0.3% agar YPD 35x10 mm plates that were made the previous day. Plates were then sealed with parafilm and incubated upright at 25°C for 10 days before they were imaged and scored for complexity (Reynolds and Fink, 2001; Hope and Dunham, 2014). Each strain was plated in duplicate on LD plates; for YPD, samples 1-62 were plated in duplicate and samples 63-78 plated in singleton.

Mat formation was scored independently by the two researchers using a four point scale; the metric was similar to that described by Hope and Dunham (2014), where 1 is indicative of a

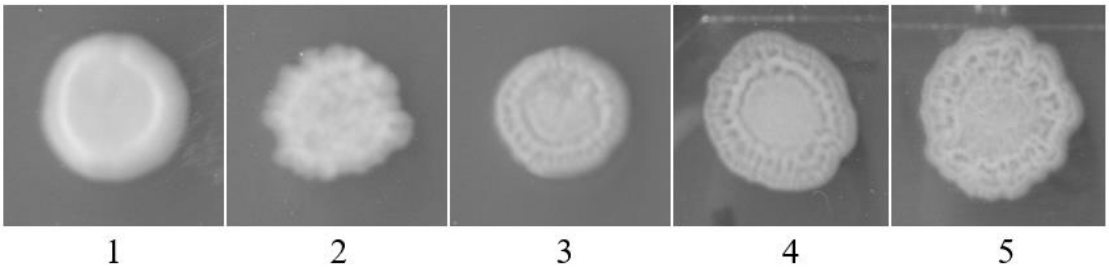
small, smooth mat, and 4 is indicative of a large, highly wrinkly mat. Example mats are provided for each value on the scale.



Complex colony formation

A 96-pin multi-blot replicator (V&P Scientific no. VP408FP6) was used to spot the 78 samples to 4 replicate OmniTrays containing solid LD medium. The plates were sealed with parafilm and incubated at 30°C for 6 days before the colonies were imaged.

The complexity of the colonies was scored similarly to the mats, but researchers used a five point scale. Here, a score of 1 is indicative of a smooth colony and a score of 5 is indicative of a complex colony formation.



Invasive growth

A 96-pin multi-blot replicator (V&P Scientific no. VP408FP6) was used to spot the 78 samples to 4 replicate synthetic low-ammonium–dextrose (SLAD) OmniTrays which were sealed with parafilm and incubated at 30°C. After three days of growth, colonies were flooded with deionized water while being gently rubbed with an index finger. Cells growing on top of the agar were washed off the plate, leaving behind only the cells that were able to grow invasively into the agar. The SLAD plates were then imaged on an Epson Expression 11000 XL scanner and strains were scored for their ability to grow invasively.

DNA Sequencing (data previously collected by Dr. Murphy)

For each of the 78 strains, genomic DNA was extracted from 5ml of overnight YPD culture using the IBI Blood and Tissue Kit. The *FLO11* locus, which included ~3kb upstream

and ~1kb downstream of the coding region, was PCR amplified using the following primers: Flo11A-for (TTGGTCAATCAGAACAGGCAAC; Chromosome IX: 388602-388624) and Flo11A-rev (GAGACATCTTTAGAGTAACCACAGATATTC; Chromosome IX: 397271-397300). Reactions were performed with iProof high-fidelity polymerase (BioRad) using the manufacturer's recommendations for cycling conditions, including adding DMSO to 3% final concentration. After inspection on an agarose gel, PCR products were cleaned using the column-based IBI PCR fragment clean-up kit. All samples were sent to the University of Georgia Genomics and Bioinformatics core for high-throughput KAPA library prep with a 96-well plate format and paired-end 300 bp sequencing on an Illumina MiSeq platform.

Assembly and Alignment

Raw reads were processed to generate *de novo* assemblies using Geneious 10.0.9R10 with the following options: Geneious Assembler, Medium Sensitivity/Fast; Re-trim Sequences. A test run with the High Sensitivity option took ~75 hours to complete, and the results did not differ from the Medium Sensitivity result, so the Medium Sensitivity was used for the rest of the samples. Contigs and consensus sequences were then exported and the longest consensus sequences for each strain were aligned to the reference *S. cerevisiae* genome with BLASTN to verify that the contigs mapped to the *FLO11* locus. Following verification, for each sample, reads were mapped back to the longest contigs in Geneious, which allowed some ambiguous sites to be resolved; consensus sequences were saved.

For most samples, the pipeline was unable to resolve the repetitive B-domain; as such, it generated separate contigs for the unique sequence of the upstream region + A-Domain and C-Domain + downstream region. The contigs were trimmed and processed to create four separate files for analysis: (1) the upstream regulatory region, (2) the downstream regulatory region, (3) the A-domain, and (4) the C-domain. Next, raw reads were mapped back to the aligned, trimmed consensus sequences with BWA (Li and Durbin, 2009) and SNPs were called with Freebayes (Garrison and Marth, 2012). High frequency SNPs were replaced in the consensus sequence file with a custom Python script. Raw reads were remapped with BWA a final time so that ambiguous sites and unexpected SNPs could be manually curated in Integrative Genomics Viewer (IGV).

In order to verify the accuracy of the pipeline, alleles from 7 strains were compared to their recently published high-resolution genomes, which are publicly available from the Yeast Population Reference Panel (YRPR; https://yjx1217.github.io/Yeast_PacBio_2016/data/) (Yue et al., 2017). The exact match between the alleles in YPRP with those obtained in this study ruled out technical artefacts as a potential contributor to the variation we observed between strains.

Characterization of B-Domain Length Variation

To characterize the length variation of the stalk-like region of the protein, the B-domain was PCR amplified using the following primers: FLO11-Repeat-for (GGTTTCGCTTGGACTGGTTGAACATGGAAC; Chromosome IX: 390616-390645) and FLO11-Repeat-rev (GATTTCCCAGGCTTCTATTGGAACATAGAT; Chromosome IX: 393076-393105). Reactions were performed with Phusion high-fidelity polymerase (New England Biolabs) using the manufacturer's recommendations for cycling conditions, including adding DMSO to 3% final concentration. Fragment analysis was conducted by running the PCR product from each sample through the Agilent 2100 BioAnalyzer using the Agilent DNA 7500 kit.

Phylogenetic Analysis

Trees were generated using PhyML in Geneious, using the HKY85 substitution model and 1000 bootstraps. Trees were generated for the A-domain alignment, the A and C-domain concatenated alignment, and the alignment of SNPs from the set of genes with similar properties to *FLO11* (see below *Comparison to Similar Genes*).

Evolutionary Analysis

The PAML (Phylogenetic Analysis by Maximum Likelihood) program package (Yang, 1997; Yang, 2007) was used to discern the patterns of selection occurring on *FLO11* (Yang and Swanson, 2002). Using codeml, both fixed sites and random sites maximum likelihood models of amino acid substitutions were used to determine the selection pressures occurring at different sites throughout the gene using various phylogenetic trees. We also analyzed the data with HyPhy (Pond *et* al., 2005) to generate the normalized dN-dS estimate. Tajima's D and π were calculated using the PopGenome (Pfeifer et al. 2014) package in R.

Comparison to Similar Genes

A literature search was conducted to identify genes with similar properties to *FLO11*, including those located within subtelomeres, encoding other cell-wall associated proteins, or encoding GPI-anchored proteins (Yue *et al*., 2017; Hamada *et al*., 1998; Klis *et al*., 2006; de Groot *et al*., 2003). A total of 135 genes were identified (listed in Table 4). Using a custom pipeline (generated by Danting Jiang), the open reading frames of these genes were extracted from the publicly available genomes (Strope *et al*., 2015, Liti *et al.,* 2009, Yue *et al*., 2017) of 54 of the 78 analyzed strains (listed in Table 5). Each gene was aligned using MAFFT (Katoh *et al*., 2002)), and using yn00 from the PAML package, the pairwise non-synonymous (dN) and synonymous rates (dS) were estimated. The alignment of the curated *FLO11* data was also analyzed with yn00 for comparison.

Data Visualization

All figures were generated in R using ggplot2 (Wikham, 2016); the phylogenetic tree was visualized using the ggtree package (Yu, et al., 2017).

Results

Phenotypic Variation

Our assays for the various social phenotypes revealed a large amount of variation between strains. For the 78 strains in our panel, some exhibited strong phenotypes for each of the three social behaviors that were examined, while many did not display much sociality. Among the three phenotypes, strains showed the greatest variation in mat formation: many strains were able to form mats moderately well, but few were highly successful at forming mats (Figure A2).

The majority of the strains did not exhibit complex colony formation, but some displayed highly complex morphologies (Figure A3). A similar pattern held for invasive growth, where the majority of strains did not exhibit strong phenotypes, but a subset of the strains were able to display invasive growth and a few were highly invasive (Figure A4). A distribution of the scores for each of these phenotypes is shown in Figure 6.

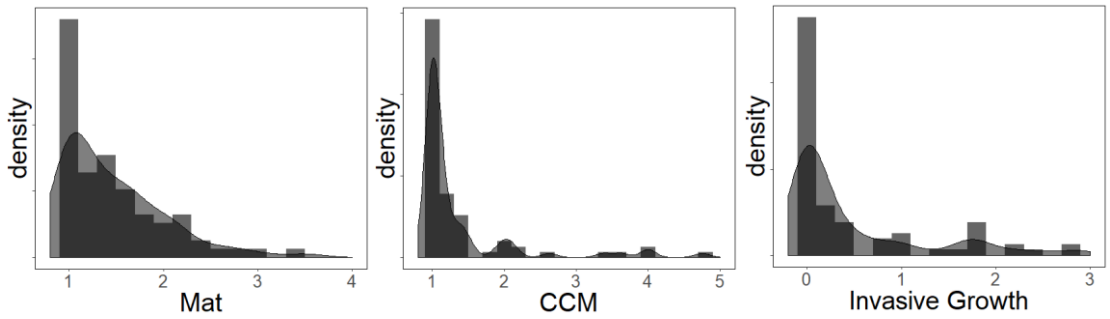


Figure 6: Distributions of scores for mat formation, complex colony morphology (CCM), and invasive growth.

Genetic Variation in the 78 Strains

Population Genetic Variation

The *de novo* assembly pipeline was able to recover alleles from 70 strains for the A domain, 76 strains for the C domain, 77 strains for the upstream region, and 78 strains for the downstream region. The alignment of *FLO11* revealed a large amount of genetic variation within both the coding and non-coding regions of the locus; there was a surprising amount of variation in the A domain (Figure A1). Close inspection of the portions of the gene coding for the A-domain revealed a large accumulation of non-synonymous mutations within the apical portions of the protein (Figure 7, Figure A5).

Evolutionary Analysis: Coding Region

The rates of nucleotide substitutions at non-synonymous sites (dN) and synonymous sites (dS) can be used to detect genes that are undergoing selection (Kimura, 1980; Kryazhimskiy and Plotkin, 2008). During the processes of transcription and translation, sets of 3 nucleotides, called codons, specify the amino acids that are included in the protein product. There are many more codons than there are amino acids; therefore, multiple codons specify the same amino acid. As a result, some nucleotide substitutions can alter codons without altering the amino acid included in the final protein product; these are called synonymous substitutions. Other nucleotide substitutions alter codons in such a way that changes the amino acid sequence of a protein product and are called non-synonymous substitutions. The ratio of these two substitution rates is often useful in elucidating the evolutionary pressures occurring on a particular gene. For example, an abundance of non-synonymous substitutions in comparison to synonymous substitutions may be indicative of an evolutionary pressure to change the protein product; this is

referred to as diversifying or positive selection. On the other hand, a paucity of non-synonymous substitutions in comparison to synonymous substitutions may be indicative of a strong selective pressure for the protein sequence to stay the same, which is called purifying selection.

It is difficult to detect positive selection by simply averaging substitution rates over all codons in a coding sequence; many amino acids may be under either structural or functional constraints, so the likelihood of having an average dN higher than an average dS is very low (Yang and Swanson, 2002). Despite the functional constraints of many codons, others may be under positive/diversifying selection (e.g., residues of the antigen-recognition site in the major histocompatibility complex). To address this issue, models have been developed which assume that there are several classes of codons in a single gene, which can be characterized by different dN/dS ratios. Some of these models have been developed for the case in which prior information is known about which sites belong to different classes and are known as fixed sites models. Other models assume that there are various site classes, but there is no *a priori* information about how to partition the sites. These models are known as random-sites models. The *FLO11* data were analyzed using codeml from PAML (Yang, 2007), a program that is able to implement these different types of models. In order to estimate dN and dS, an evolutionary model of substitutions was estimated along the branches of the phylogeny of the gene. The substitution model took into account branch lengths, codon and base frequencies and transition and transversion rates. Table 2 shows the results of the evolutionary analysis.

To test whether selection occurring within the apical region of the protein was truly different than the rest of the protein, fixed sites and random sites models were used to model the evolution of the protein in PAML. For the fixed sites models, the *FLO11* sequence was partitioned based on whether codons fell within the coding regions of the two apical regions associated with cell-cell interactions or not (Figure 7 B-C). The two models that we used, Model C and Model E in PAML, both make the assumption that the two partitions have proportional branch lengths and identical codon frequencies, but Model C assumes that the two site classes share the same dN/dS ratio, whereas Model E assumes a different dN/dS ratio for the two site classes. Comparison of these two models suggested that Model E fit the data significantly better than Model C, $\chi^2(2) = 41.526$, $p<0.001$.

Four random sites models were also used for analysis: M1, M2, M7, and M8 (Yang and Swanson, 2002). M1 is a model of neutral evolution, and assumes that there are two site classes in the sequence: conserved sites under purifying selection that have a dN/dS ratio of 0 and neutral sites that have a dN/dS ratio of 1. M2 expands on M1, adding a third site class with a free dN/dS ratio to be estimated from the data, which allows for the possibility of positive selection. Evaluating the model fit based on log-likelihood ratio values suggested that M2 fit the data significantly better than the neutral model, $\chi^2(2) = 250.794$, $p < 0.001$. We then modeled the data with M7, which assumes a beta distribution of the dN/dS ratio over all sites; the distribution is limited to the interval (0,1). M8 expands on M7 by adding another site class to M7, which can be

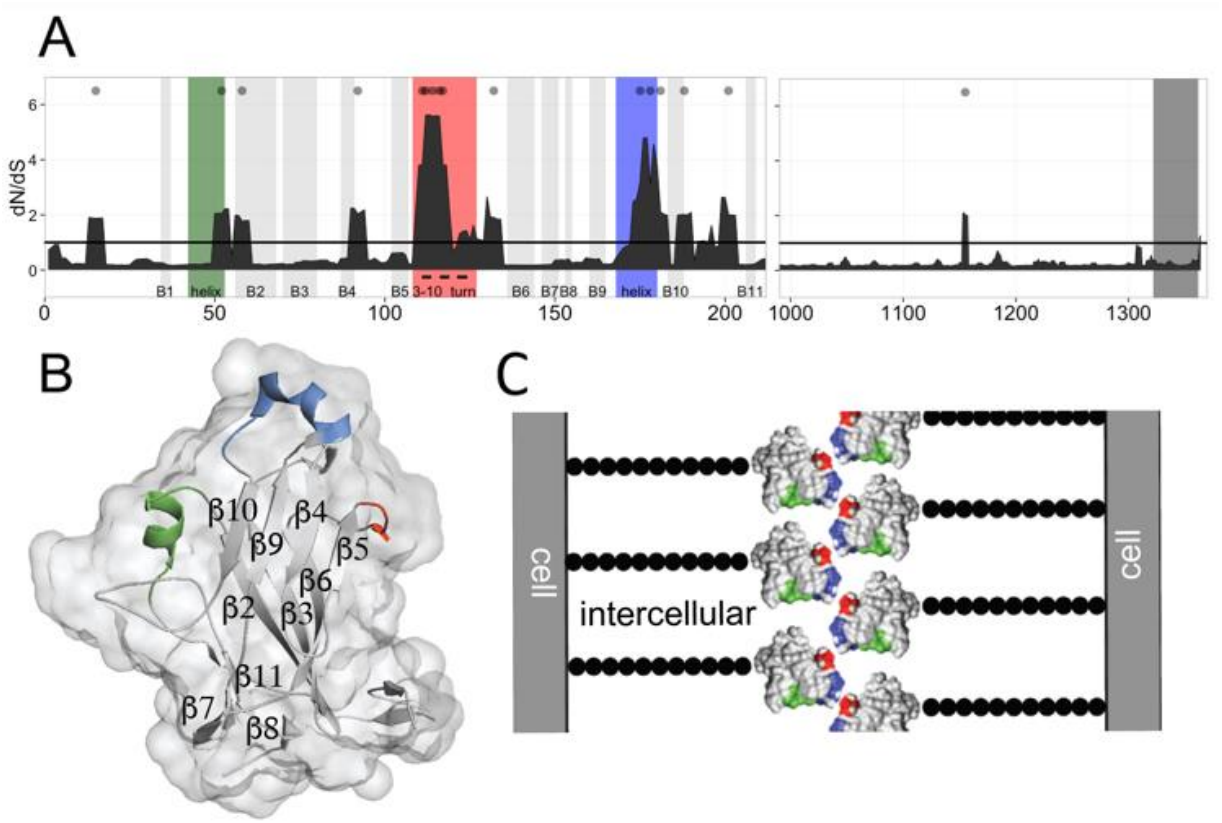


Figure 7: (A) dN/dS analysis of the concatenated A- and C-domains using a sliding window of 5 codons. Bayes Empirical Bayes postmean dN/dS values from Model 8 (see Results for details) are plotted. Points indicate sites under significant positive selection. Colored areas correspond to apical portions of the protein in the A-domain. **(B)** Structure of the A-domain, where the portions of the protein that comprise the fungal-specific apical region - the alpha helix, 3_{10} alpha helix, and the 3_{10} turn – are shown in green, blue, and red, respectively. **(C)** Schematic of a cell-cell interaction (not to scale). Domain A is hypothesized to mediate binding through the blue and red highlighted features of the apical region. Black beads represent the stalk of the protein.

estimated from the data, and allows for the possibility of positive selection. Comparison of these models suggested that M8 fit the data significantly better than M7, $\chi^2(2) = 284.204$, $p < 0.001$.

This result indicates that the models that allow for the possibility of positive selection are significantly better fits than those that do not. The models with a site class that allowed for positive selection (M2 and M8) identified codons that were under significant positive selection (Table 3). Only a single codon in the C-domain was identified, and most of the codons in the A domain were restricted to the regions in the apical portion associated with cell-cell interactions (Figure 7A-B).

Model results were robust to the phylogenetic tree that was used; trees were built either from an alignment of the A-domain, an alignment of the concatenated A- and C-domains, or an alignment of a set of 42 other genes encoding GPI or cell-wall associated proteins. The data reported above are from the models that were run using a tree that was generated from the concatenated A- and C-domain alignment; results that were obtained using other phylogenetic trees can be found in Table 2.

Evolutionary Analysis: Regulatory Regions

In order to examine the regulatory genetic variation at the *FLO11* locus, we conducted a sliding window analysis of π and Tajima's D in both the upstream and downstream regulatory regions. Pi is a measure of the mean number of pairwise differences between all of the sequences in an alignment and Tajima's D is a statistic that compares π to the total number of segregating sites in the same alignment (Tajima, 1989; Nei, 1987). This test can reveal whether a locus is undergoing neutral evolution or if it is experiencing some other selective pressure. Negative values of Tajima's D imply that there is less genetic variation than would be expected based on the number of segregating sites, which is usually a signal of purifying selection or a recent bottleneck. Positive values of Tajima's D imply that there is more genetic variation than would be expected based on the number of segregating sites, which could imply that selection is maintaining variation, a potential signal of balancing selection.

While the majority of the sites within these two regions exhibited negative Tajima's D values, one region in particular exhibited a positive Tajima's D (Figure 8). This occurred at the overlap of the two long non-coding RNAs in the upstream regulatory region. The histone

deacetylase complex, Rpd3L, binds at this site and is hypothesized to play an important role in the transitions between social and non-social phenotypes (Bumgarner *et al*., 2009).

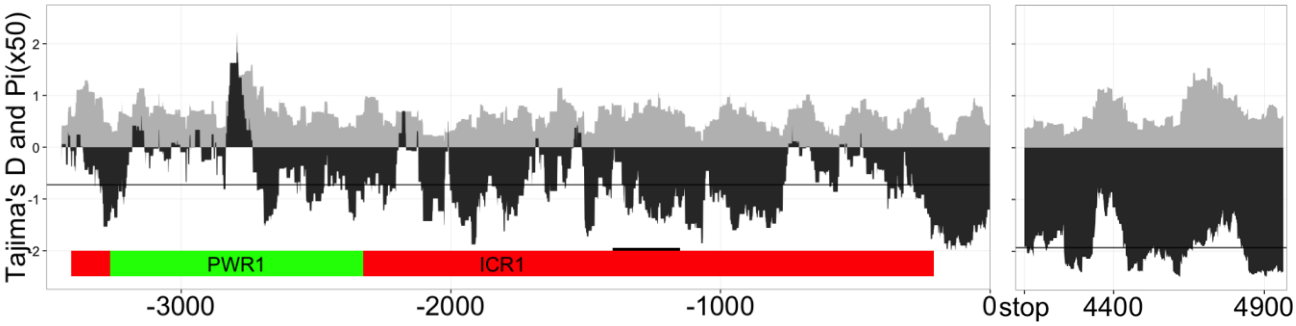


Figure 8: Sliding window analysis of Pi (in gray) and Tajima's D (in black) in the upstream regulatory region of *FLO11*. Black line indicates regional average of Tajima's D. Positive Tajima's D around -2800 bp may indicate balancing selection in this region.

B-domain Length Variation

Analysis of the B-domain confirmed that there is a large amount of length variation in the region coding for the stalk (Figure 9). The length of repetitive B domain was successfully estimated for approximately 80% of the samples. Data were not obtained for the remaining samples due either to PCR failure or an insufficient signal on the Bioanalyzer. The shortest sequence was estimated to be ~650 bp and the longest ~3.1 kb, corresponding to a stalk of ~215 and ~1000 amino acids, respectively.

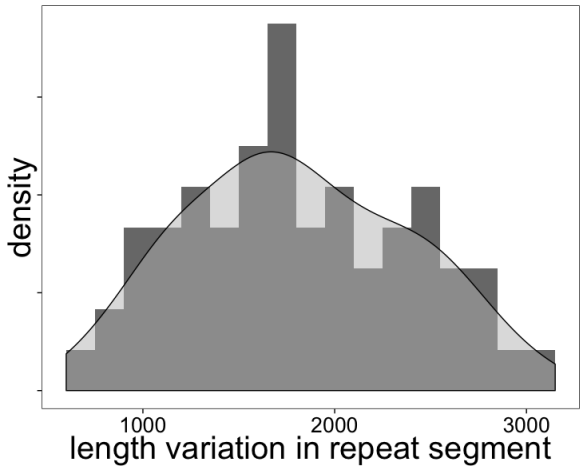


Figure 9: With a mean of 1807 bp and a standard deviation of 581, the repeat-rich region is extremely variable in length.

Phylogenetic Analysis

The phylogenetic analysis produced a tree that showed some association with ecological niche (Figure 10).

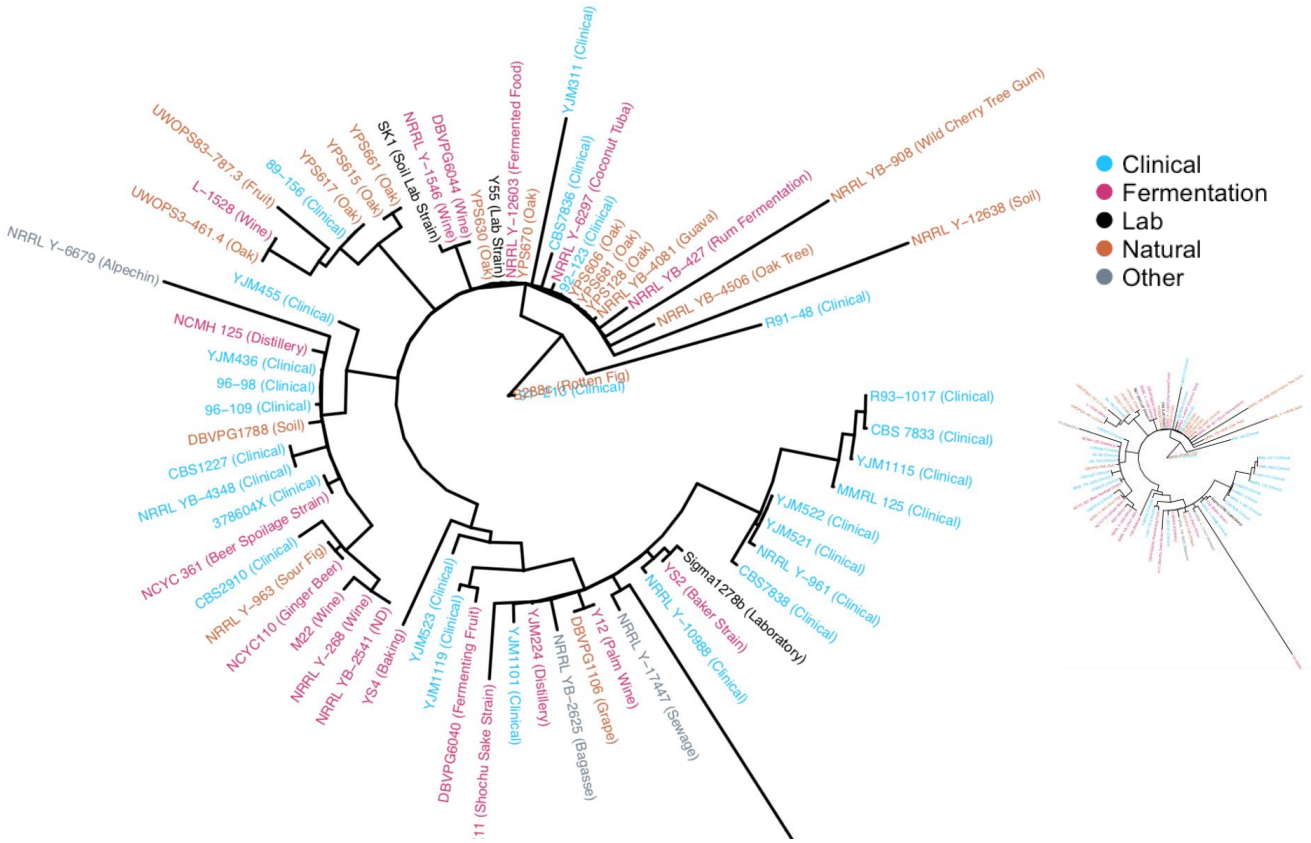


Figure 10: Phylogenetic tree generated with PhyML in Geneious and plotted using ggtree. Strains were categorized according to ecological niche, but no strong patterns of association emerged. Inset shows complete tree with Y9, a Ragi wine strain from Japan, highly diverged from the rest.

Comparison to Similar Genes

The comparison of dN and dS values between *FLO11* and other similar genes (Figure 11) revealed that *FLO11* did not have a higher rate of substitutions than other genes, although it appeared to have one of the higher ratios of the two values. Subtelomeric genes were dropped from the analysis due to poor alignment quality. It is possible genes in this category would have had dN/dS values greater than 1.

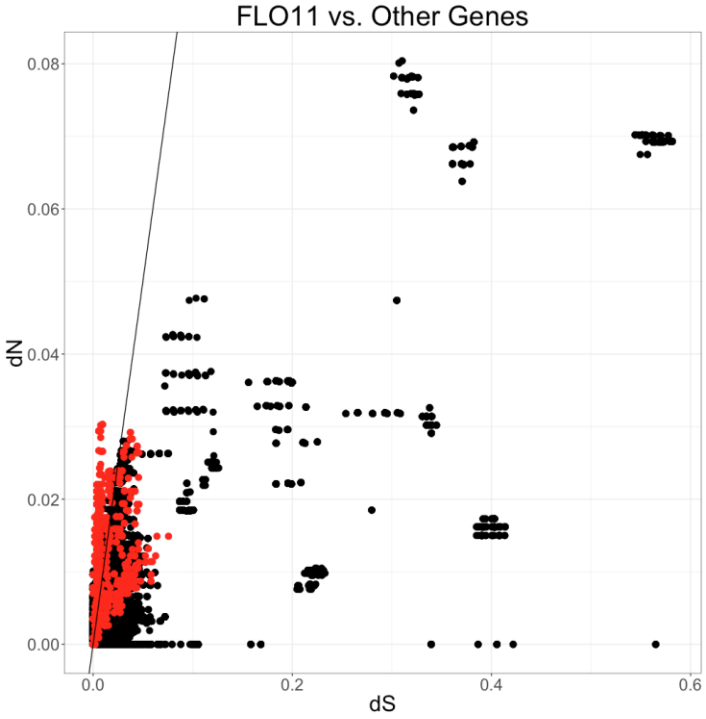


Figure 11: Pairwise dN and dS between 54 strains for *FLO11* (shown in red) and 42 other genes with similar properties (shown in black).

Testing the Functional Effect of Natural FLO11 Alleles

In order to demonstrate the functional effect of natural *FLO11* alleles on mat formation, two highly social strains were investigated. First, strains that had either one or both copies of *FLO11* knocked out were phenotyped. Strains with two *FLO11* alleles formed highly complex mats, whereas the strains with just a single allele showed a reduction in mat complexity. Furthermore, strains with no *FLO11* alleles were only able to form simple mats, demonstrating the role of *FLO11* in complex mat formation. An absence of Flo11p precludes the type of cooperation that is required to be able to form complex colonies.

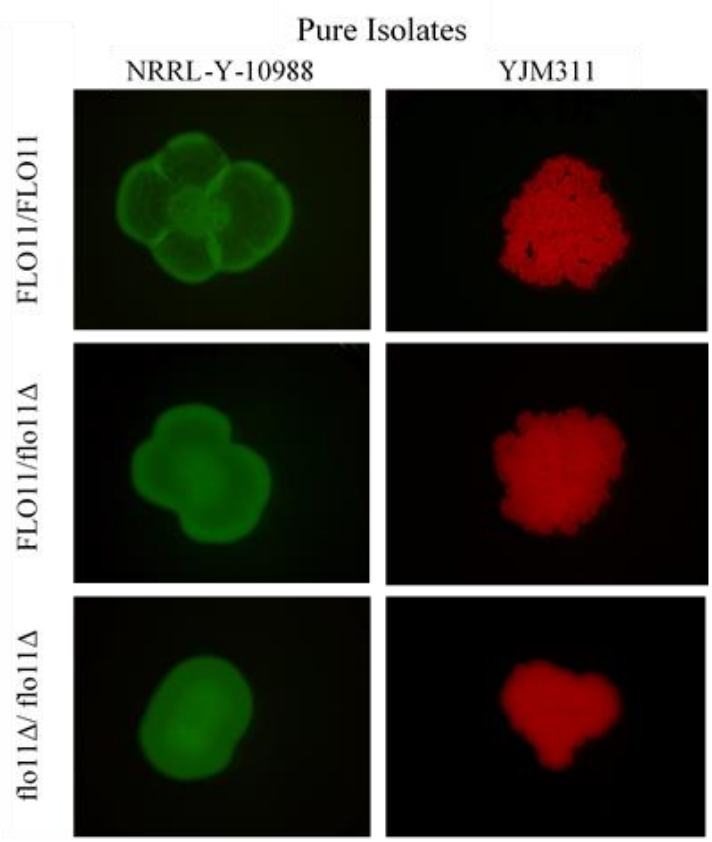


Figure 12: Two strains that exhibited strong mat formation were transformed to have either one or no copies of their natural *FLO11* allele. Strains with *FLO11* knocked out showed a reduced ability to form complex mats

The functional effect of *FLO11* was further investigated by using hemizygous hybrid strains (Steinmetz, et al., 2002) that were generated in the lab from 3 natural isolates that exhibited a range of social phenotypes (Figure 13). Multi-strains colonies in which both strains possessed the same natural allele were able to form well-mixed colonies. We found that when two strains with different alleles were plated with each other, they were able to form a single mat, but the colonies were not particularly well mixed. The segregating pattern may serve as evidence of Flo11p being a mediator of kin-recognition.

In multi-strain colonies in which multiple alleles were present, one of the alleles was typically able to out-compete the others. This occurred even when it was plated in a 1:10 ratio with another allele.

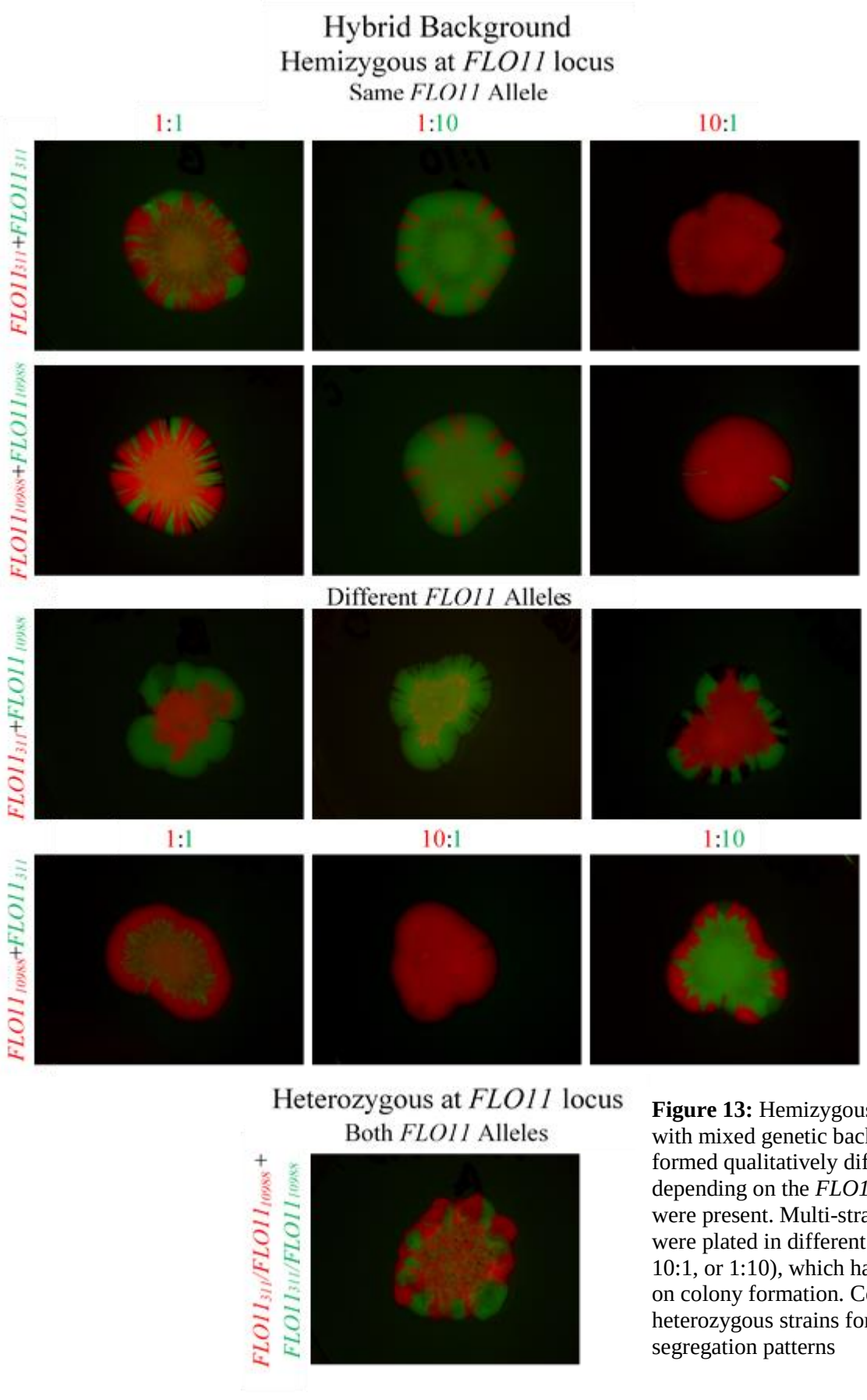


Figure 13: Hemizygous lab strains with mixed genetic backgrounds formed qualitatively different mats depending on the *FLO11* alleles that were present. Multi-strain colonies were plated in different ratios (1:1, 10:1, or 1:10), which had an impact on colony formation. Colonies of heterozygous strains formed unique segregation patterns

Discussion

The research presented here provides a compelling case for how coding and regulatory variation in the yeast cell-surface adhesin *FLO11* may provide fitness benefits through both intraclonal cooperation as well as interclonal competition. This work examined the previously uncharacterized genetic variation at the *FLO11* locus, as well as the various social phenotypes and colony forming abilities of strains with different natural *FLO11* alleles. We hypothesized that *FLO11* may serve as a kin-recognition mechanism through which individuals could engage in complex fitness-enhancing social behaviors, while avoiding the risk of cheater invasion. While we did find genetic evidence that may be suggestive of recognition-like properties of *FLO11*, phenotyping experiments revealed the importance of regulatory variation in competition between strains.

FLO11 has previously been implicated as an important mediator of social phenotypes, including those involving cell-cell adhesion, like mat and biofilm formation, as well as others, like invasive and pseudohyphal growth and plastic adherence (Guo *et al*., 2000; Lo and Dranginis, 1998; Reynolds and Fink, 2001; Verstrepen *et al*., 2004). Isolates from a wide variety of ecological niches are shown here to be capable of expressing a diverse array of social phenotypes, which until recently had gone relatively unnoticed (Liti, 2015). Strains exhibited a wide range of invasive growth ability, as well as complexity in mat formation on several different types of media. It is unclear whether the limited number of conditions assayed in the laboratory setting are a reasonable proxy for yeast social ability in a natural environment. As such, it is uncertain whether the strains which did not express strong social phenotypes in the lab are also incapable of exhibiting such behaviors in nature. Indeed, the overwhelming presence of spatially-structured, social communities in nearly all microbial species studied thus far suggests that it is unlikely that these yeast are truly planktonic in the natural environment. While social behaviors have been studied extensively in the lab over the last decade or so, there is still a gap in our knowledge regarding exactly what social behaviors are expressed by yeast in nature.

Strains possessing different *FLO11* alleles, which can be induced to express social phenotypes in the lab, are clearly expressing different levels of sociality, but it is not immediately obvious whether this difference is caused by the version of Flo11p or by the difference in complex regulatory circuitry between strains. Population genetic analysis uncovered a large amount of genetic variation across the entire locus that had previously been

uncharacterized. An in-depth evolutionary analysis revealed that while the locus as a whole is experiencing purifying selection, the distribution of nonsynonymous and synonymous substitutions within the A-domain provided evidence for positive selection; both the fixed sites and random sites models indicate that the sites coding for the apical portion of the protein are experiencing different selective pressures than other sites in the gene. These models were able to identify individual codons that are likely experiencing positive selection, and a large portion of them localized to the apical region of the A-domain. This pattern of diversifying selection in the region of the protein which allows cells to bind to one another is a pattern that is consistent with kin and self-recognition (Benabentos *et al*., 2009).

A comparison of dN and dS values between *FLO11* and other genes with similar characteristics revealed that *FLO11* does not have exceptionally elevated rates of substitution. This result is in line with the outcome of the evolutionary analysis which suggest that the *FLO11* locus in its entirety is experiencing purifying selection and only a select subset of sites are experiencing elevated rates of substitution.

Although we were unsuccessful in resolving the sequence of the B-domain, we were able to characterize the variety of stalk lengths in our panel of isolates. This result confirmed a previous finding that *FLO11* is highly polymorphic in length, although it is not clear how this length variation affects biofilm forming ability or colony morphology (Zara *et al*., 2009). We suspect that stalks of intermediate length may confer the strongest adherence, but this should be tested empirically.

Analysis of the upstream regulatory region uncovered a signal of balancing selection within the region where two long non-coding RNAs, *ICR1* and *PWR1*, overlap. The site was previously found to be a binding site for the histone deacetylase complex, Rpd3l, which is hypothesized to control the transition between social and non-social phenotypes, driving variegated expression (Bumgarner *et al*., 2009). It is interesting that genetic variation in the promoter localizes to this region, as variegated expression has been found to be an important factor in colony morphology and resource utilization (Regenberg *et al*., 2016).Variation at this site may therefore contribute to a variety of the social phenotypes observed in our panel of isolates.

We now return to our hypothesis that *FLO11* confers a fitness advantage by acting as a kin-recognition mechanism. Kin and self-recognition are not necessarily expected to evolve in

spatially structured communities, as growth generates patches of identical clones (Xavier and Foster, 2007). In such situations, the selective pressure to recognize kin is much weaker than in well-mixed populations (Strassman, et al., 2011). Thus, in order to elucidate the role of *FLO11* in self-recognition, we examined the segregating patterns of multi-allele colonies using strains with the same genetic background. Contrary to our expectations, one allele was consistently more competitive than the others, allowing the strain to take over the biofilm, even when the colony began with the "winning allele" at a numeric disadvantage of 1:10.

Taking these results in conjunction with other results from our lab, which show the existence of competition between natural isolates in spatially structured communities and a strong competitive advantage conferred by biofilms (Deschaine, et al., *in press*), we now hypothesize that ecological competition between genetic backgrounds may be a driving force in the evolution of this recognition system. Groups of clones that are better able to bind to one another to form spatially structured colonies that can efficiently utilize resources and protect themselves against environmental hazards will likely be able to outcompete other groups of clones that are less successful at forming such structures. The population genetic analysis suggests that there is a strong selective pressure on both the regulatory and coding regions of *FLO11* to mediate cooperation between clonemates in order to enable the formation of such colonies, and consequently mediate successful competition against other clones.

References

Allen, Benjamin, Jeff Gore, and Martin A. Nowak. "Spatial dilemmas of diffusible public goods." *Elife* 2 (2013).

Barrales, Ramon R., Juan Jimenez, and José I. Ibeas. "Identification of novel activation mechanisms for FLO11 regulation in Saccharomyces cerevisiae." *Genetics* 178.1 (2008): 145-156.

Barua, Subit, et al. "Molecular Basis for Strain Variation in the Saccharomyces cerevisiae Adhesin Flo11p." *mSphere* 1.4 (2016): e00129-16.

Bumgarner, Stacie L., et al. "Single-cell analysis reveals that noncoding RNAs contribute to clonal heterogeneity by modulating transcription factor recruitment." *Molecular Cell* 45.4 (2012): 470-482.

Bumgarner, Stacie L., et al. "Toggle involving cis-interfering noncoding RNAs controls variegated gene expression in yeast." *Proceedings of the National Academy of Sciences* 106.43 (2009): 18321-18326.

Crespi, Bernard J. "The evolution of social behavior in microorganisms." *Trends in Ecology & Evolution* 16.4 (2001): 178-183.

Dawkins R. 1976. The Selfish Gene. Oxford: Oxford Univ. Press

de Groot, Piet WJ, Klaas J. Hellingwerf, and Frans M. Klis. "Genome-wide identification of fungal GPI proteins." *Yeast* 20.9 (2003): 781-796.

Deschaine, B.M., Heysel, A.R., Lenhart, B.A., and H.A. Murphy. "Biofilm formation and toxin production provide a fitness advantage in mixed colonies of environmental yeast isolates." *Ecology and Evolution* (*in press).*

Douglas, Lois M., et al. "Expression and characterization of the flocculin Flo11/Muc1, a Saccharomyces cerevisiae mannoprotein with homotypic properties of adhesion." *Eukaryotic Cell* 6.12 (2007): 2214-2221.

Dranginis, Anne M., et al. "A biochemical guide to yeast adhesins: glycoproteins for social and antisocial occasions." *Microbiology and Molecular Biology Reviews* 71.2 (2007): 282-294.

Fidalgo, Manuel, et al. "Adaptive evolution by mutations in the FLO11 gene." *Proceedings of the National Academy of Sciences* 103.30 (2006): 11228-11233.

Fidalgo, Manuel, Ramon R. Barrales, and Juan Jimenez. "Coding repeat instability in the FLO11 gene of Saccharomyces yeasts." *Yeast* 25.12 (2008): 879-889.

Frieman, Matthew B., and Brendan P. Cormack. "Multiple sequence signals determine the distribution of glycosylphosphatidylinositol proteins between the plasma membrane and cell wall in Saccharomyces cerevisiae." *Microbiology* 150.10 (2004): 3105-3114.

Garrison, Erik, and Gabor Marth. "Haplotype-based variant detection from short-read sequencing." *arXiv preprint arXiv:1207.3907* (2012).

Goossens, Katty VY, and Ronnie G. Willaert. "The N-terminal domain of the Flo11 protein from Saccharomyces cerevisiae is an adhesin without mannose-binding activity." *FEMS Yeast Research* 12.1 (2012): 78-87.

Gullerova, M., and N. J. Proudfoot. "Transcriptional Interference and Gene Orientation in Yeast Noncoding RNA Connections." *Cold Spring Harbor Symposia on Quantitative Biology*. Cold Spring Harbor Laboratory Press, 2011.

Guo, Bing, et al. "A Saccharomyces gene family involved in invasive growth, cell–cell adhesion, and mating." *Proceedings of the National Academy of Sciences* 97.22 (2000): 12158-12163.

Halme, Adrian, et al. "Genetic and epigenetic regulation of the FLO gene family generates cell-surface variation in yeast." *Cell* 116.3 (2004): 405-415.

Hamada, K., et al. "Screening for glycosylphosphatidylinositol (GPI)-dependent cell wall proteins in Saccharomyces cerevisiae." *Molecular and General Genetics MGG* 258.1-2 (1998): 53-59.

Hamada, K., et al. "Amino acid residues in the ω-minus region participate in cellular localization of yeast glycosylphosphatidylinositol-attached proteins." *Journal of Bacteriology* 181.13 (1999): 3886-3889.

Hamilton, William D. "The genetical evolution of social behaviour. II." *Journal of Theoretical Biology* 7.1 (1964): 17-52.

Hope, Elyse A., and Maitreya J. Dunham. "Ploidy-regulated variation in biofilm-related phenotypes in natural isolates of Saccharomyces cerevisiae." *G3: Genes, Genomes, Genetics* 4.9 (2014): 1773-1786.

Huang, Guohong, Mingliang Zhang, and Scott E. Erdman. "Posttranslational modifications required for cell surface localization and function of the fungal adhesin Aga1p." *Eukaryotic Cell* 2.5 (2003): 1099-1114.

Katoh, Kazutaka, et al. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." *Nucleic Acids Research* 30.14 (2002): 3059-3066.

Keller, Laurent, and Michael G. Surette. "Communication in bacteria: An ecological and evolutionary perspective." *Nature Reviews Microbiology* 4.4 (2006): 249.

Kimura, Motoo. "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." *Journal of Molecular Evolution* 16.2 (1980): 111-120.

Klis, Frans M., Andre Boorsma, and Piet WJ de Groot. "Cell wall construction in Saccharomyces cerevisiae." *Yeast* 23.3 (2006): 185-202.

Kraushaar, Timo, et al. "Interactions by the fungal Flo11 adhesin depend on a fibronectin type III-like adhesin domain girdled by aromatic bands." *Structure* 23.6 (2015): 1005-1017.

Kuthan, Martin, et al. "Domestication of wild Saccharomyces cerevisiae is accompanied by changes in gene expression and colony morphology." *Molecular Microbiology* 47.3 (2003): 745-754.

Legras, J. L., et al. "Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history." *Molecular Ecology* 16.10 (2007): 2091.

Li, H. and R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 2009. **25**: p. 1754-60

Liti, Gianni. "The Natural History of Model Organisms: The fascinating and secret wild life of the budding yeast S. cerevisiae." *Elife* 4 (2015): e05835.

Lo, Wan-Sheng, and A. M. Dranginis. "FLO11, a yeast gene related to the STA genes, encodes a novel cell surface flocculin." *Journal of Bacteriology* 178.24 (1996): 7144-7151.

Lo, Wan-Sheng, and A. M. Dranginis. "The cell surface flocculin Flo11 is required for pseudohyphae formation and invasion by Saccharomyces cerevisiae." *Molecular Biology of the Cell* 9.1 (1998): 161-171.

Nadell, Carey D., Knut Drescher, and Kevin R. Foster. "Spatial structure, cooperation and competition in biofilms." *Nature Reviews Microbiology* 14.9 (2016): 589.

Nather, Kerstin, and Carol A. Munro. "Generating cell surface diversity in Candida albicans and other fungal pathogens." *FEMS Microbiology Letters* 285.2 (2008): 137-145.

Nayyar, Ashima, et al. “Flocculation in industrial strains of Saccharomyces cerevisiae: role of cell wall polysaccharides and lectin-like receptors.” *Journal of the Institute of Brewing* 123.2 (2017): 211-218.

Nei, Masatoshi. *Molecular evolutionary genetics*. Columbia university press, 1987.

Nett, Jeniel E., and David Andes. "Fungal biofilms: in vivo models for discovery of anti-biofilm drugs." *Microbiology Spectrum* 3.3 (2015): E30.

Octavio, Leah M., Kamil Gedeon, and Narendra Maheshri. "Epigenetic and conventional regulation is distributed among activators of FLO11 allowing tuning of population-level heterogeneity in its expression." *PLoS Genetics* 5.10 (2009): e1000673.

Palková, Z. (2004). Multicellular microorganisms: Laboratory versus nature. *EMBO Reports*, *5*(5), 470-476.

Paulick, Margot G., and Carolyn R. Bertozzi. "The glycosylphosphatidylinositol anchor: a complex membrane-anchoring structure for proteins." *Biochemistry* 47.27 (2008): 6991-7000.

Pfeifer, B. et al. (2014) PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Mol Biol Evol* 31(7): 1929-1936., doi:10.1093/molbev/msu136

Pond, Sergei L. Kosakovsky, and Spencer V. Muse. "HyPhy: hypothesis testing using phylogenies." *Statistical Methods in Molecular Evolution*. Springer, New York, NY, 2005. 125-181.

Regenberg, Birgitte, et al. "Clonal yeast biofilms can reap competitive advantages through cell differentiation without being obligatorily multicellular." *Proc. R. Soc. B*. Vol. 283. No. 1842. The Royal Society, 2016.

Reynolds, Todd B., and Gerald R. Fink. "Bakers' yeast, a model for fungal biofilm formation." *Science* 291.5505 (2001): 878-881.

Rupp, Steffen, et al. "MAP kinase and cAMP filamentation signaling pathways converge on the unusually large promoter of the yeast FLO11 gene." *The EMBO Journal* 18.5 (1999): 1257-1269.

Smith, J. Maynard. "Group selection and kin selection." *Nature* 201.4924 (1964): 1145.

Soares, Eduardo V. "Flocculation in Saccharomyces cerevisiae: a review." *Journal of Applied Microbiology* 110.1 (2011): 1-18.

Steinmetz, L.M., H. Sinha, D.R. Richards, J.I. Spiegelman, P.J. Oefner, J.H. McCusker, et al., "Dissecting the architecture of a quantitative trait locus in yeast." *Nature*, **416**(6878): 326-30 (2002).

Strassmann, Joan E., Owen M. Gilbert, and David C. Queller. "Kin discrimination and cooperation in microbes." *Annual Review of Microbiology* 65 (2011): 349-367.

Strope, Pooja K., et al. "The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen." *Genome Research* 25.5 (2015): 762-774.

Tajima, Fumio. "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." *Genetics* 123.3 (1989): 585-595.

Váchová, Libuše, et al. "Flo11p, drug efflux pumps, and the extracellular matrix cooperate to form biofilm yeast colonies." *The Journal of Cell Biology* (2011): jcb-201103129.

Van Dyken, J. David, et al. "Spatial population expansion promotes the evolution of cooperation in an experimental prisoner's dilemma." *Current Biology* 23.10 (2013): 919-923.

Van Mulders, Sebastiaan E., et al. "Flocculation gene variability in industrial brewer's yeast strains." *Applied Microbiology and Biotechnology* 88.6 (2010): 1321-1331.

Van Mulders, Sebastiaan E., et al. "Phenotypic diversity of Flo protein family-mediated adhesion in Saccharomyces cerevisiae." *FEMS Yeast Research* 9.2 (2009): 178-190.

Verstrepen, Kevin J., and Frans M. Klis. "Flocculation, adhesion and biofilm formation in yeasts." *Molecular Microbiology* 60.1 (2006): 5-15.

Verstrepen, Kevin J., Todd B. Reynolds, and Gerald R. Fink. "Origins of variation in the fungal cell surface." *Nature Reviews Microbiology* 2.7 (2004): 533-540.

Voordeckers, Karin, et al. "Identification of a complex genetic network underlying Saccharomyces cerevisiae colony morphology." *Molecular Microbiology* 86.1 (2012): 225-239.

West, Stuart A., et al. "Social evolution theory for microorganisms." *Nature Reviews Microbiology* 4.8 (2006): 597-607.

West, Stuart A., et al. "The social lives of microbes." *Annu. Rev. Ecol. Evol. Syst.* 38 (2007): 53-77.

Wickham, Hadley. *ggplot2: elegant graphics for data analysis*. Springer, 2016.

Xavier, J.B. and K.R. Foster, "Cooperation and conflict in microbial biofilms." *Proceedings of the National Academy of Sciences* (2007), **104**(3): p. 876-881.

Yang, Ziheng. "PAML: a program package for phylogenetic analysis by maximum likelihood." *Bioinformatics* 13.5 (1997): 555-556.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution, 24*(8), 1586-1591. doi:10.1093/molbev/msm088

Yang, Ziheng, and Willie J. Swanson. "Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes." *Molecular Biology and Evolution* 19.1 (2002): 49-57.

Yu G, Smith D, Zhu H, Guan Y and Lam TT (2017). “ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data.” *Methods in Ecology and Evolution*, **8**, pp. 28-36.

Yue, Jia-Xing, et al. "Contrasting evolutionary genome dynamics between domesticated and wild yeasts." *Nature Genetics* 49.6 (2017): 913.

Zara, Giacomo, et al. "FLO11 gene length and transcriptional level affect biofilm-forming ability of wild flor strains of Saccharomyces cerevisiae." *Microbiology* 155.12 (2009): 3838-3846.

Table 1. List of the 78 strains, along with their parental isolate name, geographic and environmental origin, as well as population membership (according to ref 1.). The data obtained for each strain is indicated with a check.

Sample	HMY	Strain	Parental isolate/ isogenic with	Isolate source or other name	Geographic origin	Origin	Population	Upstream	Domain A	Domain C	Downstream	B Length	Reference
1	2		YJM224		Unknown	Distillery		✔	✔	✔	✔	✔	
2	3		YJM311	91-212	CA, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	
4	25	YJM244	YJM210	CBS 1227	Romania	Clinical	Wine/Eur	✔	✔	✔	✔	✔	1
3	26	YJM248	YJM218	CBS 2910	Unknown	Clinical (feces)	Wine/Eur	✔	✔	✔	✔	✔	1
5	27	YJM270	YJM215	CBS 2807	Slovenia	Wine	Wine/Eur	X	X	X	X	X	1
6	30	YJM326	YJM310	CBS 7838	CA, USA	Clinical	Mosaic	✔	✔	✔	✔	X	1
7	31	YJM428	YJM308	CBS 7836	CA, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1
8	32	YJM450	YJM440	92-123	CA, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1
9	33	YJM451	YJM436	B70302(b)	Unknown	Clinical	Mosaic	✔	✔	✔	✔	✔	1
10	35	YJM456	YJM454	89-156	CA, USA	Clinical	Mosaic	✔	✔	✔	✔	X	1
11	36	YJM470	YJM455	SUH	CA, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1
12	37	YJM541	YJM522	SUH	CA, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1
13	38	YJM554	YJM521	SUH	CA, USA	Clinical	Mosaic	✔	✔	✔	✔	X	1
14	39	YJM555	YJM523	SUH	CA, USA	Clinical	Mosaic	✔	✔	✔	✔	X	1
15	43	YJM693	YJM669	R91-48	TX, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1
16	45	YJM969	YJM967	96-98	Italy	Clinical	Wine/Eur	✔	✔	✔	✔	✔	1
17	46	YJM689	YJM675	R93-1017	TX, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1
18	47	YJM972	YJM947	96-100	Italy	Clinical	Wine/Eur	✔	X	X	✔	X	1
19	57	YJM1078	YJM1075	NRRL YB-4348	Portugal	Clinical	Wine/Eur	✔	✔	✔	✔	✔	1
20	58	YJM1083	YJM1073	NRRL Y-10,988	NC, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1

21	59	YJM1129	YJM1123	NCMH 125	OH, USA	Distillery	Wine/Eur	✔	✔	✔	✔	✔	1
22	60	YJM1133	YJM1125	MMRL 125	NC, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1
23	62	YJM1199	YJM1115		MI, USA	Clinical	Mosaic	✔	✔	✔	✔	X	1
24	64	YJM1208	YJM1119	1882	MI, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1
25	67	YJM1248	YJM1219	NRRL Y-1546	West Africa	Wine	West African	✔	✔	✔	✔	✔	1
26	69	YJM1252	YJM1224	NRRL Y-6679	Spain	Alpechin	Wine/Eur	✔	✔	✔	✔	✔	1
27	72	YJM1290		Sigma1278b	Unknown	Laboratory	Mosaic	✔	✔	✔	✔	X	1
28	75	YJM1307	YJM1071	NRRL Y-961	DC, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1
29	76	YJM1311	YJM1101	C. Kaufman	MI, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1
30	80	YJM1338	YJM1315	NRRL Y-963	MD, USA	Sour fig	Mosaic	✔	✔	✔	✔	✔	1
31	82	YJM1342	YJM1324	NRRL Y-12638	South Africa	Soil	Mosaic	✔	✔	✔	✔	✔	1
32	85	YJM1381	YJM1357	NRRL YB-427	Trinidad	Rum fermentation	Mosaic	✔	✔	✔	✔	✔	1
33	88	YJM1386	YJM1365	NRRL Y-11878	Jamaica	Sugar cane	Mosaic	✔	✔	X	✔	✔	1
34	91	YJM1389	YJM1368	NRRL Y-17447	Thailand	Sewage	Sake	✔	✔	✔	✔	✔	1
35	92	YJM1399	YJM1393	NRRL YB-908	Unknown	Wild cherry tree gum	Mosaic	✔	✔	✔	✔	✔	1
36	93	YJM1400	YJM1394	NRRL YB-4081	Philippines	Guava	Mosaic	✔	✔	✔	✔	✔	1
37	96	YJM1415	YJM1407	NRRL Y-268	France	Wine	Wine/Eur	✔	✔	✔	✔	✔	1
38	97	YJM1417	YJM1411	NRRL YB-2541	HI, USA	ND	Wine/Eur	✔	✔	✔	✔	✔	1
39	98	YJM1418	YJM1413	NRRL YB-4506	Japan	Oak tree	Mosaic	✔	✔	✔	✔	✔	1
40	99	YJM1419	YJM1412	NRRL YB-2625	Unknown	Bagasse	Mosaic	✔	✔	✔	✔	✔	1
41	100	YJM1433		Yllc17_E5	France	Wine	Wine/Eur	✔	X	X	✔	X	1,2
42	103	YJM1443		UWOPS83-787.3	Bahamas	Fruit	Mosaic	✔	✔	✔	✔	✔	1,2
43	111	YJM1479	YJM1474	NRRL Y-6297	Phillipines	Coconut tuba	Mosaic	✔	✔	✔	✔	✔	1
44	117	YJM1615	YJM312	91-213	CA, USA	Clinical	Mosaic	✔	✔	✔	✔	✔	1
45	121	YJM1573	YJM1566	NRRL Y-12603	India	Fermented food	Mosaic	✔	✔	✔	✔	✔	1
46	122	YJM1552	S1	S288c	CA, USA	Rotten fig	Mosaic	✔	✔	✔	✔	✔	1
47	124	YJM1529		M22	Unknown	Wine	Wine/Eur	✔	✔	✔	✔	✔	1
48	158			YPS128	PA, USA	Oak	North American	✔	✔	✔	✔	✔	3

50	160			L-1528	Chile	Wine	Wine/Eur	✔	✔	✔	✔	X	2
49	161	YJM975	YJM948	96-101	Italy	Clinical	Wine/Eur	✔	X	X	✔	✔	1
51	169			SK1	USA	Soil, lab strain	West African	X	✔	✔	✔	✔	2
52	170			K11	Japan	Sake strain	Sake	✔	✔	✔	✔	✔	2
53	171			UWOPS05-217.3	Malaysia	Bertam Palm	Malaysian	✔	X	X	✔	✔	2
54	172			NCYC361	Ireland- *S. diastaticus*	Beer spoilage	Mosaic	✔	✔	✔	✔	✔	2
55	173			DBVPG6765	unknown- *S. boulardi*	Unknown	Wine/Eur	✔	X	X	✔	X	2
56	174	YJM1549		DBVPG6040	Netherlands- *S. fructum*	Fermenting fruit juice	Mosaic	✔	✔	✔	✔	✔	1,2
57	175	YJM1463		DBVPG1853	Ethiopia	White tecc	Wine/Eur	✔	X	X	✔	✔	1,2
58	217	YJM145	YJM128	CBS 7833	MO, USA	Clinical		✔	✔	✔	✔	X	1
59	261	YJM978	YJM954	96-109	Italy	Clinical	Wine/Eur	✔	✔	✔	✔	✔	1
60	262	YJM1434		YPS606	PA, USA	Oak	North American	✔	✔	✔	✔	✔	3
61	263			L-1374	Chile	Wine	Wine/Eur	✔	X	X	✔	✔	2
62	264	YJM1460		Y12	Africa	Palm Wine	Sake	✔	✔	✔	✔	✔	1,2
63	265			YS2	Australia	Baker strain	Mosaic	✔	✔	✔	✔	✔	2
64	266	YJM1439		NCYC110	West Africa	Ginger beer	West African	✔	✔	✔	✔	X	1,2
65	267			DBVPG1106	Australia	Grape	Wine/Eur	✔	✔	✔	✔	✔	2
66	268			378604X	Newcastle, UK	Clinical	Mosaic	✔	✔	✔	✔	✔	2
67	270			YPS681	PA, USA	Oak	North American	✔	✔	✔	✔	✔	4
68	271			YPS615	PA, USA	Oak	North American	✔	✔	✔	✔	✔	4
69	272			YPS617	PA, USA	Oak	North American	✔	✔	✔	✔	✔	4
70	273			YPS661	PA, USA	Oak	North American	✔	✔	✔	✔	✔	4
71	274			YPS670	PA, USA	Oak	North American	✔	✔	✔	✔	✔	4
72	275			YPS630	PA, USA	Oak	North American	✔	✔	✔	✔	✔	4
73	276			UWOPS3-461.4	Malaysia	Bertam Palm, nectar	Malaysian	✔	✔	✔	X	✔	2
74	277			Y55	France	Wine, lab strain		X	✔	✔	✔	✔	2

75	278		YS4	Netherlands	Baker strain	Mosaic	✔	✔	✔	✔	✔	2
76	279		DBVPG1788	Finland	Soil	Wine/Eur	✔	✔	✔	✔	✔	2
77	280		DBVPG6044	West Africa	Bili wine	West African	✔	✔	✔	✔	✔	2
78	281		Y9; NRRL-Y5997	Japan	Ragi wine	Sake	✔	✔	✔	✔	X	2

References

1. Strope, P.K., D.A. Skelly, S.G. Kozmin, G. Mahadevan, E.A. Stone, P.M. Magwene, et al. (2015), The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. **Genome Research**, 25(5): p. 762-774.

2. Liti, G., D.M. Carter, A.M. Moses, J. Warringer, L. Parts, S.A. James, et al. (2009), Population genomics of domestic and wild yeasts. **Nature**, 458: p. 337-341.

3. Sniegowski, P.D., P.G. Dombrowski and E. Fingerman (2002), *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. **FEMS Yeast Research**, 1: p. 299-306.

4. Murphy, H.A. and C.W. Zeyl (2012), Prezygotic isolation between *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* through differences in mating speed and germination timing. **Evolution**, 66(4): p. 1196-1209

Table 2: Results of the evolutionary analysis conducted in PAML using three different phylogenetic trees. For each tree, random sites models which allowed for the possibility of positive selection (2 and 8) were significantly better fits than those that not (1 and 7). Likewise, the fixed sites model that allowed the dN/dS value to differ between the site partitions (E) was a significantly better model of the data than the model that kept dN/dS constant between the partitions (C).

Tree	Model	Log Likelihood	Parameters	AIC	Comparison	Test Statistic	P-value
A/C Concatenated	Model1	-6626.686	111	13475.4	2 vs. 1	250.794	2.20E-16
	Model2	-6501.289	113	13228.6			
	Model7	-6626.727	112	13477.5	8 vs. 7	284.204	2.20E-16
	Model8	-6484.625	114	13197.3			
	ModelC	-6724.781	121	13691.6	E vs. C	41.526	9.61E-10
	ModelE	-6704.018	123	13654.0			
A-domain	Model1	-2059.398	103	4324.8	2 vs. 1	49.106	2.17E-11
	Model2	-2034.845	105	4279.7			
	Model7	-2059.504	104	4327.0	8 vs. 7	48.966	2.33E-11
	Model8	-2035.021	106	4282.0			
	ModelC	-2100.346	113	4426.7	E vs. C	39.858	2.21E-09
	ModelE	-2080.417	115	4390.8			
Core Gene Tree	Model1	-7310.787	77	14775.6	2 vs. 1	888.794	2.20E-16
	Model2	-6866.390	79	13890.8			
	Model7	-7320.583	78	14797.2	8 vs. 7	906.836	2.20E-16
	Model8	-6867.165	80	13894.3			
	ModelC	-7527.352	87	15228.7	E vs. C	103.500	2.20E-16
	ModelE	-7475.602	89	15129.2			

Table 3: Codons were identified by models that allowed for positive selection (M2 and M8) as being under significant positive selection. Red and blue codons are associated with the apical regions of the A-domain and grey codons fall within the B-domain. Numbers refer to the amino acid within the concatenated A and C-domain alignment and numbers in parentheses refer to the amino acid within the reference strain, S288c, *FLO11* sequence.

A and C-Domain	
Model 2	17N, 54D, 60Q, 94W, 113Y, 114-, 116-, 118-, 119-, 134Y, 177Q, 180S, 183Q, 190D, 203H, [217A(852), 219V(854), 220S(855), 228V(863), 256S(891), 274Q(909), 299P(934),] 522K (1157)
Model 8	17N, 54D, 60Q, 94W, 113Y, 114-, 116-, 118-, 119-, 134Y, 177Q, 180S, 183Q, 190D, 203H, [217A(852), 219V(854), 220S(855), 228V(863), 235S(870), 236S(871), 250F(885), 253T(888), 256S(891), 257F(892), 274Q(909), 293T(928), 299P(934),] 522K(1157)
A-Domain Tree	
Model 2	4P, 6L, 17N, 54D, 57N, 94W, 113Y, 114-, 116-, 118-, 119-, 177Q, 179A, 180S, 183Q, 190D
Model 8	4P, 6L, 17N, 54D, 57N, 91K, 94W, 113Y, 114-, 116-, 118-, 119-, 124-, 125-, 126-, 130N, 134Y, 174S, 177Q, 179A, 180S, 183Q, 190D, 196N, 200N, 203H
Core Gene Tree	
Model 2	4P, 6L, 7L, 17N, 54D, 57N, 60Q, 85K, 91K, 94W, 106G, 113Y, 114-, 116-, 118-, 119-, 128-, 129-, 131E, 133T, 172Q, 173G, 175A, 176A, 177Q, 178Y, 179A, 182W, 189F, 195C, 202G, [219V(854), 227S(862), 290T(925), 298T(933), 327T(962), 334T(969), 335T(970), 338T(973), 343S(978),] 521G(1156), 550T(1185), 552A(1187), 675S(1310), 731V(1366)
Model 8	4P, 6L, 7L, 17N, 54D, 57N, 60Q, 85K, 91K, 94W, 106G, 113Y, 114-, 116-, 118-, 119-, 129-, 131E, 133T, 172Q, 173G, 175A, 176A, 177Q, 178Y, 179A, 182W, 189F, 195C, 202G, [219V(854), 227S(862), 290T(925), 298T(933), 327T(962), 334T(969), 335T(970), 338T(973), 343S(978),] 521G(1156), 550T(1185), 552A(1187), 675S(1310), 731V(1366)

Table 4: Genes recovered from literature search for genes with similar properties to *FLO11*. Only the bolded genes were ultimately used in the dN/dS analysis; we were unable to acquire sufficiently resolved alignments for the other genes to be able to include the in the analysis.

YAL063C	**YEL040W**	**YIL011W**	YLL061W	YNL336W
YAL068C	YEL069C	YIL169C	YLL062C	**YNR044W**
YAR050W	YEL070W	YIL172C	YLL064C	YNR072W
YAR071W	YER011W	YIL176C	YLR037C	YNR073C
YBL108C-A	**YER150W**	YIR039C	**YLR040C**	YNR076W
YBR067C	YER185W	YIR041W	**YLR042C**	YOL030W
YBR162C	YFL058W	YJL078C	**YLR110C**	**YOL132W**
YBR297W	YFL060C	**YJL158C**	**YLR194C**	YOL154W
YBR299W	YFL061W	**YJL159W**	YLR300W	YOL155C
YBR301W	YFL062W	**YJL171C**	**YLR342W**	YOL157C
YBR302C	**YGL028C**	YJL214W	**YLR343W**	YOL161C

YCL069W	**YGL228W**	YJL216C	**YLR390W-A**	**YOR009W**
YCL073C	YGL261C	YJL221C	YLR461W	**YOR010C**
YCR089W	**YGR279C**	YJL222W	YML132W	**YOR214C**
YCR096C	YGR292W	YJL223C	YMR215W	YOR388C
YCR098C	YGR294W	**YJR004C**	**YMR238W**	YOR389W
YCR102C	YGR295C	YJR156C	**YMR305C**	YOR390W
YCR104W	**YHL009C**	YJR158W	**YMR307W**	YOR391C
YCR107W	YHL044W	YJR161C	YMR325W	YOR393W
YDL243C	YHL046C	**YKL046C**	**YNL190W**	YOR394W
YDL244W	YHL048W	**YKL096W**	**YNL300W**	YPL279C
YDL245C	**YHR126C**	**YKL096W-A**	**YNL322C**	YPL280W
YDL248W	**YHR143W**	YKL219W	**YNL327W**	YPL281C
YDR055W	YHR211W	YKL224C	YNL331C	YPL282C
YDR077W	YHR213W	YKR105C	YNL332W	**YPR159W**
YDR261C	YHR215W	YKR106W	YNL334C	YPR194C
YDR349C	YHR216W	YLL025W	YNL335W	YPR196W
YDR542W				

Table 5: Strains used in dN/dS analysis for similar genes to *FLO11*

Sample	HMY	Strain		Sample	HMY	Strain
3	HMY26	CBS2910		34	HMY91	NRRL Y-17447
4	HMY25	CBS1227		35	HMY92	NRRL YB-908
5	HMY27	CBS2807		36	HMY93	NRRL YB-4081
6	HMY30	CBS7838		37	HMY96	NRRL Y-268
7	HMY31	CBS7836		38	HMY97	NRRL YB-2541
8	HMY32	92-123		39	HMY98	NRRL YB-4506
9	HMY33	YJM436		40	HMY99	NRRL YB-2625
10	HMY35	89-156		41	HMY100	YJM1433
11	HMY36	YJM455		43	HMY111	NRRL Y-6297
12	HMY37	YJM522		44	HMY117	91-213
13	HMY38	YJM521		45	HMY121	NRRL Y-12603
14	HMY39	YJM523		46	HMY122	S288c
15	HMY43	R91-48		48	HMY158	YPS128
16	HMY45	96-98		49	HMY161	96-101
17	HMY46	R93-1017		51	HMY169	SK1
18	HMY47	96-100		55	HMY173	DBVPG6765
20	HMY58	NRRL Y-10988		56	HMY174	DBVPG6040
21	HMY59	NCMH 125		57	HMY175	DBVPG1853
22	HMY60	MMRL 125		57	HMY175	DBVPG1853
23	HMY62	YJM1115		59	HMY261	96-109
24	HMY64	YJM1119		60	HMY262	YPS606
25	HMY67	NRRL Y-1546		62	HMY264	Y12
26	HMY69	NRRL Y-6679		62	HMY264	Y12
28	HMY75	NRRL Y-961		64	HMY266	NCYC110
29	HMY76	YJM1101		73	HMY276	UWOPS3-461.4
30	HMY80	NRRL Y-963		77	HMY280	DBVPG6044
32	HMY85	NRRL YB-427		-	HMY269	UWOPS83-787.3

Appendix

Figure A1: A-Domain Alignment (apical regions shaded green, red, and blue)

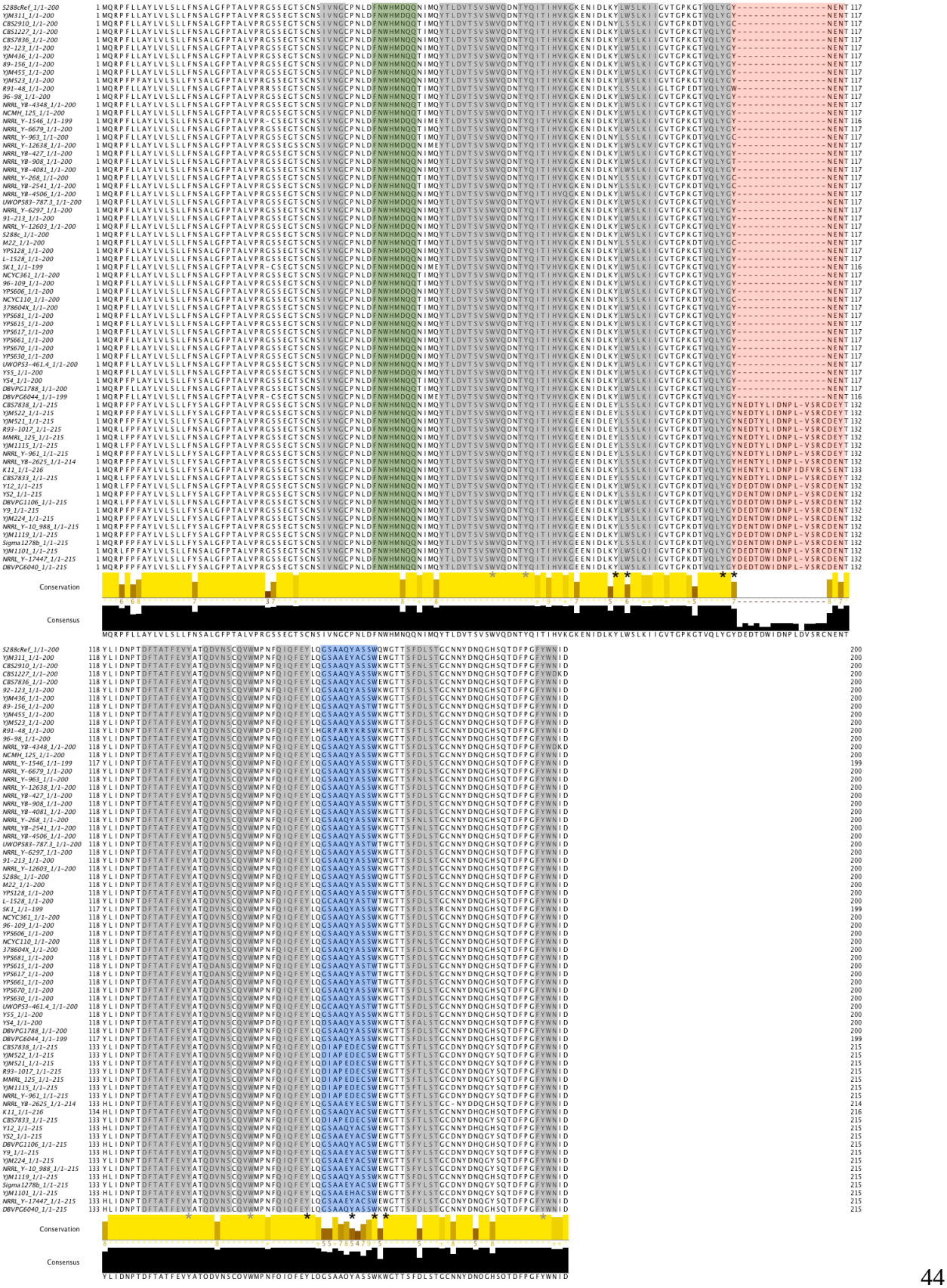


Figure A2: Mat Formation Assay for 78 Strains

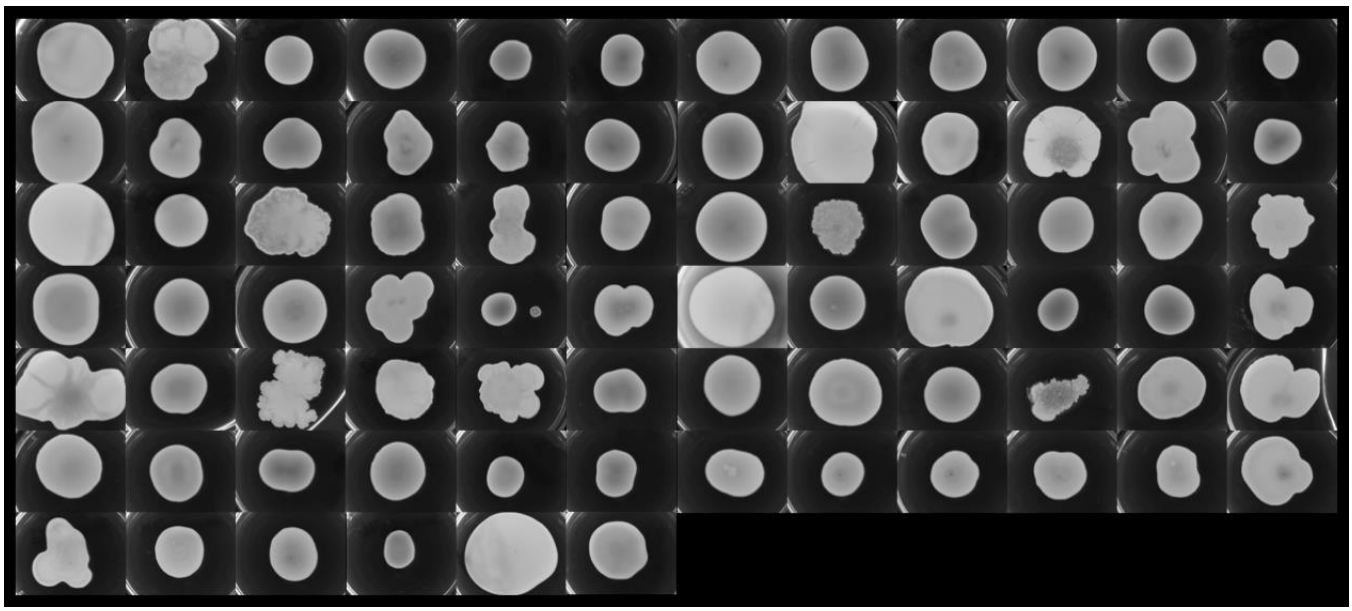


Figure A3: Complex Colony Formation Assay

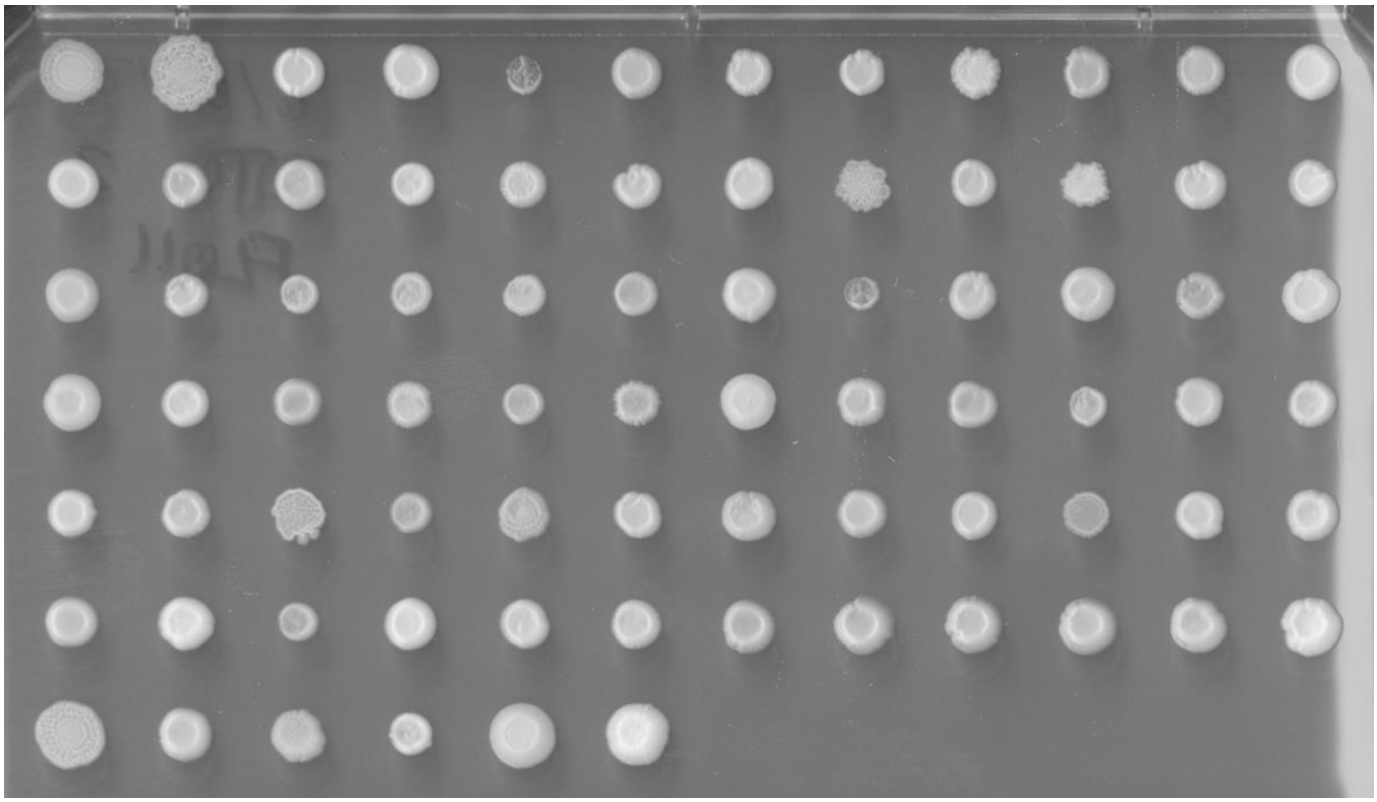


Figure A4: Invasive Growth Assay

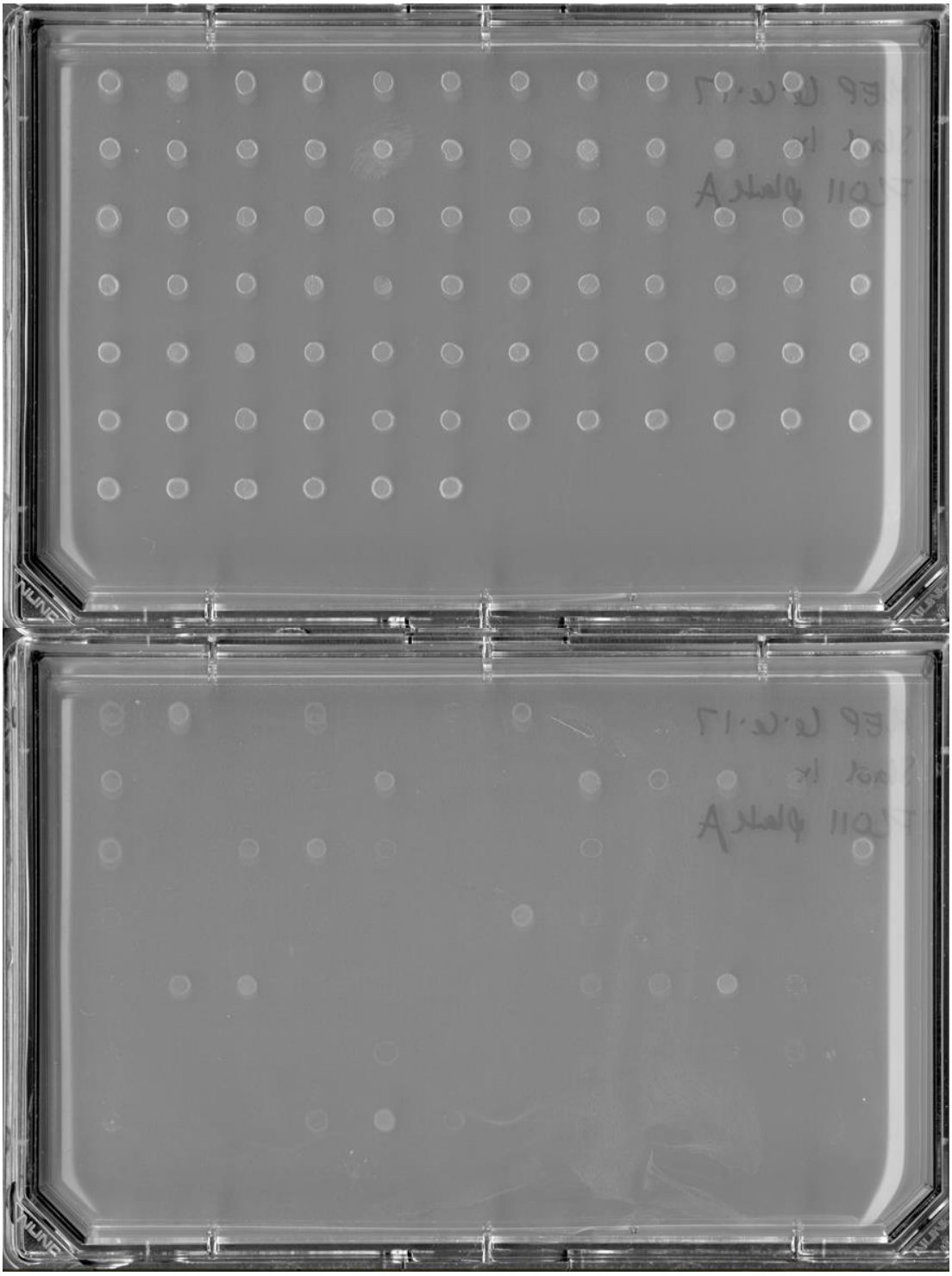
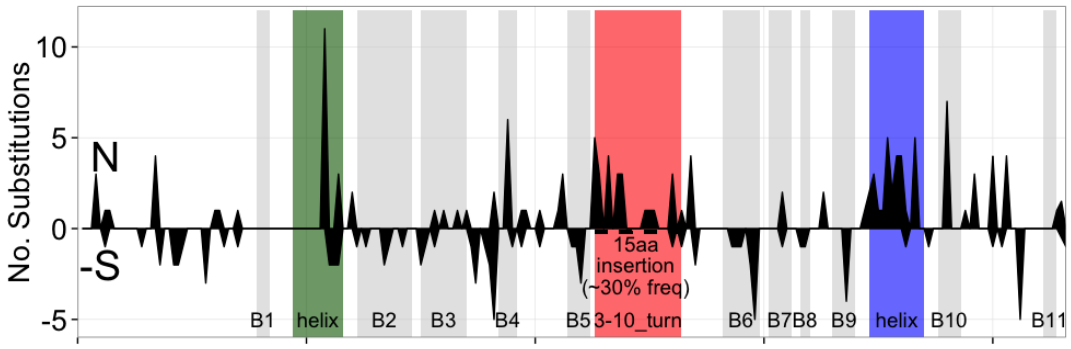


Figure A5:



A sliding window analysis in HyPhy uncovered an abundance of synonymous and nonsynonymous substitutions within the A-domain of *FLO11*. Nonsynonymous substitutions accumulated in the apical regions of the protein in particular.